Supermarket Sales

1. Business Understanding

**Client**

Supermarket chain in Karnataka (Bangalore, Mangalore and Mysore)

**Business goal**

To improve the gross income of the supermarket.

**Business questions**

1. Which city has the maximum gross income? Is there any difference in aggregate sales across all the branches?
2. Which city is paying the maximum tax?
3. How’s the unit price affected by gross income, quantity, and other variables like day, timeslot, and product line?
4. Find and explain the relationship between product line and gross income for Branch?

Find and explain the distribution of the Gross income and Unit Price.

**Business Model:**

Univariate Analysis and Bivariate Analysis

1. We have identified the target feature as Customer type.
2. Target feature is balanced class. Member and Normal are balanced.
3. We need to find Gross income of the 3 cities.
4. Which city is paying more tax.
5. How the gross income is distributed.
6. How the Unit price is distributed,
7. How the product type and Gross income distributed.

2. Ideas

3. Selected Idea

1. **Which city has the maximum gross income? Is there any difference in aggregate sales across all the branches?**

A distribution graph node was used to find the frequency of the categorical field “Branch”. There is no much of a difference in sales across all the branches A, B, and C. But, the sales in Branch C is a bit higher than the rest of the branches. Here’s the numeric data of the branch sales, **Branch**

c 5265.1765

a 5057.1605

b 5057.0320

**City**

Mysore 5265.1765

Bangalore 5057.1605

Mangalore 5057.0320

1. **How’s the unit price affected by gross income, quantity, and other variables like day, timeslot, and product line?**

4. Data Acquisition

* **How many features**

16 features are there in the dataset

* **Size of the dataset** 1000 rows are there in the dataset.
* **Multiple files**

Single Supermarket sales dataset file. Dataset contains various attributes of supermarket like Invoice Id, Branch, City, Customer Type, Gender, Product Type, Unit Price, Quantity, Tax, Selling Price, Date, Time, Payment Type, Cost Price, Gross Income, and Rating.

* **What kind of data – numerical or character** Both Numerical data and Character data are

5. Understanding

1. We have 14 features in the dataset,
2. A distribution graph node was used to find the frequency of the categorical field “Branch.
3. Bar plot used to identify the target feature whether it is balanced or imbalanced.
4. Bar plot used to identify the gross income, tax paid per city and gross income per product line.
5. Create a model to analyse the relationship between different variables of the supermarket.
6. Create a clustering model to identify how the other features are related to the target feature

#Tax, Selling and Cost Price has upper tail

outliers.

6. Wrangling 1

No Missing Value , #No NA, so no need of NA imputation.

7. Wrangling 2

1. Replacing outliers values with 90% quantile values for upper tail and 10% quantile values for lower tail outliers for tax.
2. Replacing outliers values with 90% quantile values for upper tail and 10% quantile values for lower tail outliers for cost price.

We can see that all Product\_Type are having above 12.5 Gross\_Income and below 16, Which shows all Product Type are similar in Gross\_Income

**3. Find and explain the distribution of the Gross income and Unit Price?** Gross Income and tax are Rightly skewed. Unit Price is normally distributed.

1. Standardize the

variables on the training data set.

1. Feature engineering techniques:
   1. For Categorical feature: We have used One-Hot

8. Feature Selection

present on the Input dataset.

It is evident that Invoice ID is detected as “typeless” since it has hyphens in between which could not be detected as a valid measurement type.

Whereas, fields like branch, city, product line and payment are “nominal” that have multiple values. The customer type and gender are a binary field, and thus has “flag” measurement. And the rest of the variables are “continuous” since they have a range of numeric values. We are dropping off Invoice ID and Time.

12. Results of ML technique 1

Classification Repo rt is: (**Randome For st)**

prec

ision recall f1

-score support

member

0.47 0.47

0.47 150

normal

0.47 0.47

0.47 150

accuracy

0.47 300

macro avg

0.47 0.47

0.47 300

weighted avg 0.47 0.47

0.47 300

Encoding: Which has created separate features for all levels of categorical features

1. For Numerical features: We have used StandardScaler for scaling.
2. Features extration from Date columns

10. Results of FE technique 2

Correlation between variables and dependent variable

**Conclusion:**

Tax, Selling Price,Cost Price are highly positive correlated to Target feature (Gross\_Income)

#Therefore we will drop Tax, Selling\_Price,Cost\_Price features.

Replacing outliers values with 90% quantile values for upper tail and 10% quantile values for lower tail outliers for selling price

9.Results of FE technique 1

.Year,Month and Day features extrated from Date

1. Logit is a Parametric ML Model, So before passing on numerical feature with different scales.
2. We need to normalize the feature One-Hot Encoding for categorical
3. Histogram to find how the data distributed on Gross income and Unit Price and various other features.
4. Box plot to identify the Outliers and found that the features Tax, selling price and the cost price

11. Modeling

**Random forest** is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

#One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

**Logistic Regression** is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

#Contrary to popular belief, logistic regression IS a regression model. The model

0.52 300

weighted avg 0.53 0.53

0.52 300

Confusion Matrix: [[95 55]

[86 64]]

0.53

0.53

accuracy

0.53 300

macro avg

0.48

0.43

0.54

0.63

0.52

member

150

normal

150

0.57

recall f1-

precision support

score

13. Results of ML technique 2

Classification Report is: (**Logistic reg regression)**

Model 1: Baseline RandomForest Algorithm

gives 47% accuracy

Model 2: GridSearch Tunned RandomForest Algorithm gives 49% accuracy

Model 3: Baseline LogisticRegression Algorithm gives 53% accuracy

Model 4: GridSearch Tunned LogisticRegression Algorithm gives 50% accuracy

a. From above models accuracy then can see that even with model hyper parameter tuning no good increase in accuracy has happened

14. Comparison

1. We used Uni-variate, bi-variate and

correlation analysis to perform basic EDA on the supermarket sales data.

1. To summarize below are some of the findings/observations from the data:
2. The customer rating is more or less uniform with the mean rating being around 7 and there is no relationship between gross income and customer ratings.
3. The data consists of 3 cities/branches. Though branch C has slightly higher sales than the rest, A and B.
4. There is no particular time trend that can be observed in gross income.

15. Conclusion

builds a regression model to predict the probability that a given data entry belongs to the category numbered.

16. Recommendation / Suggestion

1. We can offer Loyalty Programs so that we can attract more customers and so that they can get enrolled in the membership,
2. The supermarket's own data can be used to complete the picture, analyzing customer loyalty to competitors' products, sensitivity to category promotions and the results of past new product launches.

Confusion Matrix: [[70 80]

[80 70]]

Confusion Matrix: [[74 76]

[83 67]]

1. We need to try other feature engineering techniques for model building.
2. LogisticRegression baseline model is giving us best accuracy
3. At an overall level, ‘Sports and Travel’, home and life style generates highest gross income.
4. Using the correlation analysis, one interesting observation has emerged that customer ratings is not related to any variable.
5. Though the rating for ‘fashion accessories’ and ‘food and beverages’ is high but the quantity purchased is low. Hence, supply for these products need to be increased.