**Methodology**

* **Logistic Regression**

Logistic regression is a popular classification algorithm used to assign observations to a discrete set of classes. It’s a predictive analysis algorithm and based on the concept of probability. It transforms output using sigmoid function to return a probability value. It’s much easier to train, implement, efficient, fewer assumptions of variables and decent degree of accuracy

* **Decision Tree**

Decision Tree is used for both classification and regression problems and mostly preferred for classification problems. Its tree structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules and each lead node represents the outcome

**Overview**

**Understanding the business problem:**

A Supermarket is a self-service shop offering a wide variety of food, beverages and household products, organized into sections. The growth of supermarkets in most populated cities is increasing and market competitions are also high. The Super Market Sales Analytics is used to analyze the sales of different products in a supermarket located in different cities.

In the most populated cities, the number of supermarkets is growing, and market competition is fierce. The dataset is a record of the historical sales made by a grocery chain over a three-month period in 3 separate branches.

**What is the problem that you are trying to solve?**

How can the supermarket enhance sales by foreseeing the key characteristics that will lead to greater profit and greater customer happiness using the provided dataset?

**Dataset & Understanding of Data**

* **Size of the Dataset**

It has got 1000 data values and 16 features

* **Variable Type**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| At**tributes** | **Data Type** | **Category** | **Type** | **Description** |
| Invoice ID | Numeric | Nominal | Predictor | Computer generated sales slip invoice identification number |
| Branch | Character | Categorical | Predictor | Branch of supercentre’s (3 branches are available identified by A, B and C), |
| City | Character | Categorical | Predictor | Location of supercentre’s |
| Customer Type | Character | Categorical | Predictor | Type of customers, recorded by Members for customers using member card and Normal for without member card |
| Gender | Character | Nominal | Predictor | Gender type of customer |
| Product Type | Character | Categorical | Predictor | General item categorization groups |
| Unit Price | Numeric | Ratio | Predictor | Price of each product |
| Quantity | Numeric | Ratio | Predictor | Number of products purchased by customer |
| Tax | Numeric | Ratio | Predictor | Tax fee for customer buying |
| Selling Price | Numeric | Ratio | Predictor | Price at which product is sold |
| Date | Numeric | Continuous | Predictor | Date of purchase (Record available from January 2019 to March 2019) |
| Time | Numeric | Ratio | Predictor | Purchase time |
| Payment Type | Character | Categorical | Predictor | Payment used by customer for purchase (3 methods are available – Cash, Credit card and Debit Card) |
| Cost Price | Numeric | Ratio | Predictor | Cost of goods |
| Gross Income | Numeric | Ratio | Predictor | Sum of all the wages |
| Rating | Numeric | Ratio | Target | Customer stratification rating on their overall shopping experience (On a scale of 1 to 10) |

* **Data Distribution:**

The Supermarket Sales Dataset has 1000 rows × 16 columns. There are 9 categorical features and 6 Numerical attributes. The Supermarket Sales has balanced dataset.

**Feature Wrangling**

* **Missing Value**:

For the dataset for the supermarket, missing value treatment was handled through imputed values.

* **Irrelevant Value :**
  + Date & Time will not help in prediction so this have to be removed
  + ‘Invoice ID’ is just ID number and doesn’t help in prediction so this is removed
* **Reshaping Value :**

Reshaping the dataset – Selling Price, Cost Price, Tax, Gross Income has outliers and it’s been handled.

* **EDA outcomes and discussion**

Used univariate, bi-variate, and correlation analysis to visualize Exploratory Data Analysis on the supermarket sales data.

**Customer Type & City:**

Customer type was categorized into Normal and Members. 3 Cities are considered for analysis. That are Bangalore, Mangalore and Mysore. Customer type “Member” is slightly higher then “Normal” i.e., 50.1% belongs to Members and 49.9% belongs to Normal.

**Payment Type & City:**

Payment types are categorized into Debit, Credit card & Cash type. Debit card is the most preferred payment method in all the cities. Mysore has cash payment as higher rate compared to another city. Bangalore has more debit payment type then the other city.

**Selling Price Variation in City per Date:**

Each city's total selling price is represented visually and indexed in accordance with the timeline.

**Product Type Related Factors:**

Product type has been classified into 6 variations such as Fashion, Electronic accessories, Food and beverages, Sports and travel, Home & lifestyle, Health and beauty.

**Feature Engineering & Selection**

Categorical values are given dummy values and then embedded into final data model. So final Data set size used for modeling & prediction is (1000, 16)

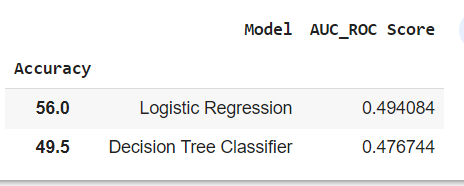
Two techniques used for feature selection are

|  |  |
| --- | --- |
| **Random Forest Classifier** | **XgBoost** |
| This is a Bagging Algorithm which aggregates a specified number of decision trees. These decrease impurities (Gini Impurity) over all trees & improve purity of then node. At start of the trees, greatest decrease in impurity will be at start of the trees and nodes with the least decrease in impurity occur at the end of trees. This will help in selecting feature of most importance. | Gradient boosting automatically provides estimates of feature importance from a trained predictive model. After boosted trees are constructed, it is easy to retrieve importance scores for each attributes. |
|  |  |
| **Feature removed**  Based on the feature engineering techniques used for importance , below features are removed from the final data set which is of very low value.   1. Weekday 2. Month 3. Gross Income | **Feature removed**  Based on the feature engineering techniques used for importance , below features are removed from the final data set which is of very low value.   1. Customer Type 2. Gender 3. Branch 4. Month 5. Product Type |

**Modeling Results**

|  |  |
| --- | --- |
| * **Logistic Regression** | * **Decision Tree Classifier** |
| precision recall f1-score support  0 0.57 0.96 0.71 114  1 0.33 0.02 0.04 86  accuracy 0.56 200  macro avg 0.45 0.49 0.38 200  weighted avg 0.47 0.56 0.43 200  AUC\_ROC Score: 0.4940840473276214  Accuracy of LGR: 56.0 % | precision recall f1-score support  0 0.55 0.50 0.52 114  1 0.41 0.45 0.43 86  accuracy 0.48 200  macro avg 0.48 0.48 0.48 200  weighted avg 0.49 0.48 0.48 200  AUC\_ROC Score: 0.4767441860465116  Accuracy of Decision Tree (CART): 48.0 % |
|  |  |

**Conclusion**



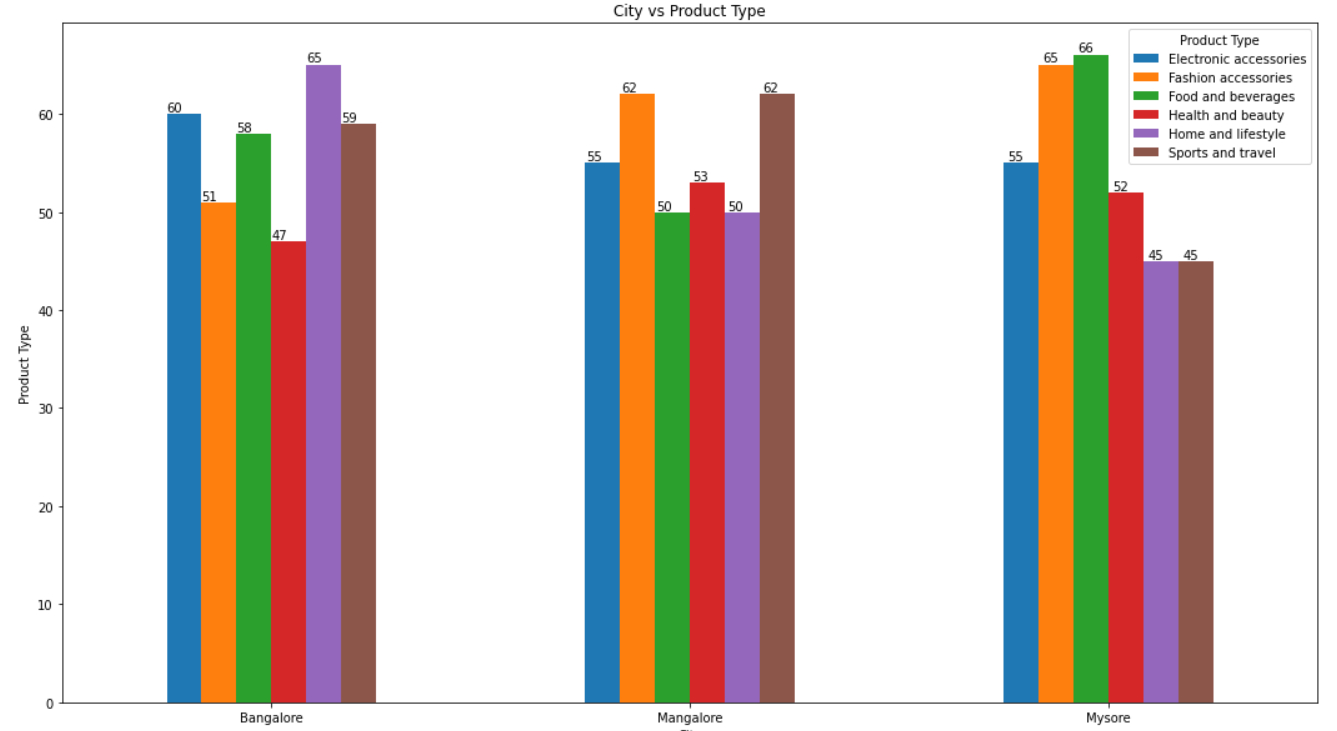
|  |  |
| --- | --- |
|  | * Above accuracy clearly tells that Logistic Regression is the best model to predict Super Market Sales Analytics * Model Synopsis: * The most accurate model is the logistic regression model, which is neither overfitted nor underfitted. * There is no correlation between gross income and customer ratings, and the average rating is consistent at approximately 7. * There are 3 cities/branches in the data. Branch C, or Mysore, is the most profitable branch in terms of gross income even though branch A has somewhat higher sales than the others. * The most popular items in Mysore are food and drink items and fashion accessories. Along with technological accessories, these things should be prioritized. * Debit cards are the most widely used form of payment. Also on the expensive side, cash payment. |

Objective

# Q1) Which city has a better sale for products in the Electronic Accessories product line.

# Answer:

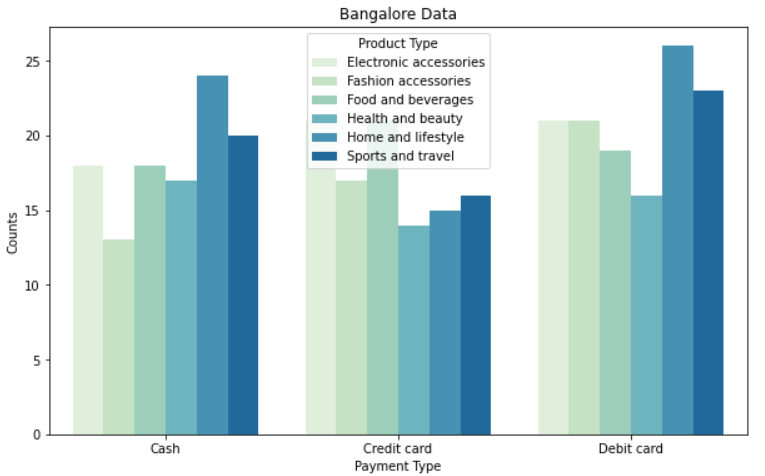
**Bangalore** city has a better sale for products in the Electronic Accessories product line.



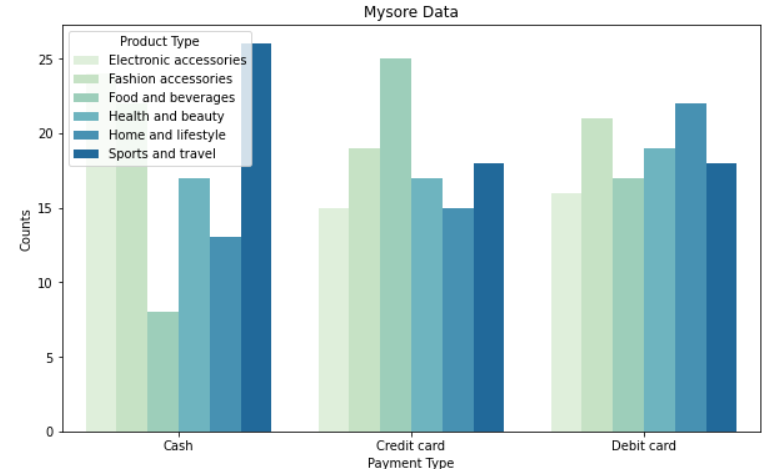
# Q2) Which payment method is used more often at a particular city, branch and for which product type

# Answer:

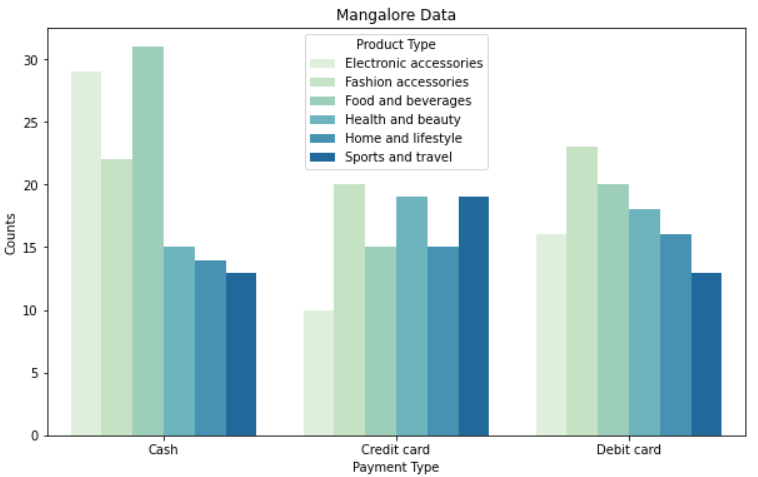
1. **Bangalore** City: As can be seen from the graph above, the product type category for which **debit cards** are used more frequently is Home & Lifestyle. Cash is the second most common payment method in Bangalore City, with a count of 110, followed by credit cards with a count of 104.



(2) **Mysore** City: As you can see from the graph above, the category of products for which **debit cards** are used more frequently is Home & Lifestyle, and the most common payment method in Bangalore is debit cards, with a count of 113, followed by cash with a count of 110, and credit cards with a count of 109.

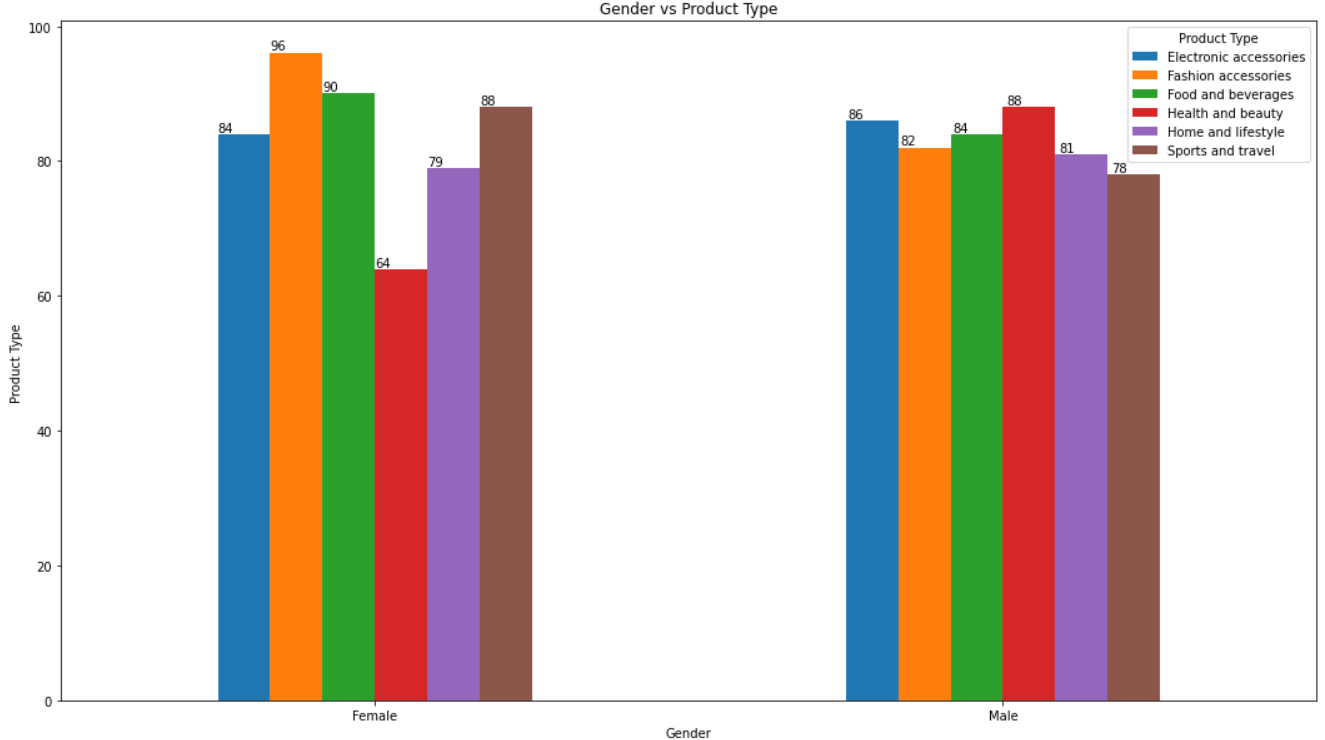


(3) **Mangalore** City: As you can see that **Cash** is the most used payment type in mangalore city with count of 124 followed by Debit card with the count 106 and lastly credit card with count 98 and the Food & Beverages is the product type category for which Debit card is used more oftenly, which is clearly shown in the graph above for mangalore city.



# Q3) Which Product type has been more purchased by female customers

# Answer:

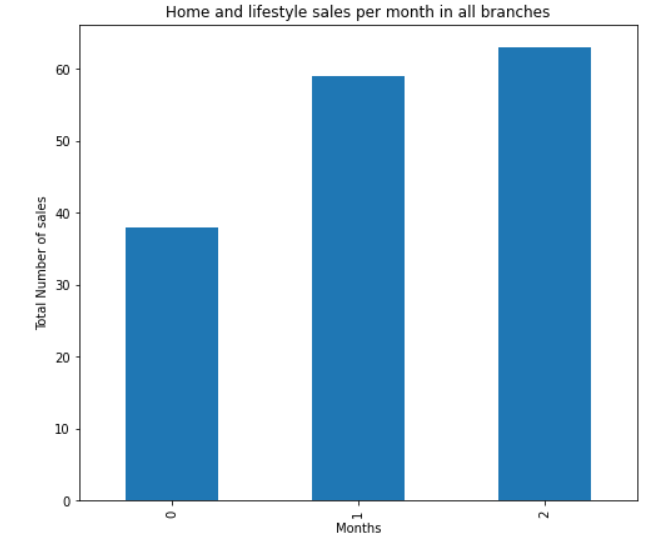
**Fashion accessories** product line has been more purchased by female customers.

# Q4) In which month does the highest number of home and lifestyle products have been sold

# Answer:

**March** month does the highest number of home and lifestyle products have been sold.

Legends: 0-Jan, 1-Feb, 2-March



# Q5) At what time most of the female customers are purchasing products.

# Answer:

At **10 am** most of the female customers are purchasing products followed by **1 pm** and then by **7 pm**

* 