**Project report on**

**Supermarket Sales prediction using Linear Regression and Random Forest Classifier**

****

**Description:**  
  
The growth of supermarkets in most populated cities are increasing

and market competitions are also high. The dataset is one of the historical sales of super market company which has been recorded in 3 different branches for 3 months data.

**Attribute information:**  
  
Invoice id: Computer generated sales slip invoice identification number  
  
Branch: Branch of super center (3 branches are available identified by A, B and C).  
  
City: Location of super centers  
  
Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card.  
  
Gender: Gender type of customer  
  
Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel  
  
Unit price: Price of each product in $  
  
Quantity: Number of products purchased by customer  
  
Tax: 5% tax fee for customer buying  
  
Total: Total price including tax  
  
Date: Date of purchase (Record available from January 2019 to March 2019)  
  
Time: Purchase time (10am to 9pm)  
  
Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and E-wallet)  
  
COGS: Cost of goods sold  
  
Gross margin percentage: Gross margin percentage  
  
Gross income: Gross income  
  
Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

Problem Statement:

In this project we will analyse the sales dataset of a supermarket and predict how it can develop in the future and thee steps to be taken for customer satisfaction.

Predicting the future supermarket sales using Linear Regression and Random Forest Classifier considering

X(dependent variable):Sum of the gross margin percentage of a particular day.

Y(independent variable): Sum of unit price, date(converted to number of days), gender, product type, city, member(Categorical variables converted to continuous).

Visualization:

## Conclusion from data visualization

## 1. The gross percent spreads with unit price, i.e for a unit price of higher value, there are a wide range of gross percentages of the commodities in sale

## 2. Branch C of the supermarket makes the greatest profit.

## 3. City of Naypyitaw draws maximum percentage of customers.

## 4. Most people have voted up to a rating of 6.0 which says that the services and facilities are good for all branches.

## 5. From the heatmap we can analyse that tax 5%, total, cogs and gross income are highly correlative.

## 6. Mode of payment doesn't matter at all.

## 7. The survey was performed on equal no. of men and women.

## 8. No. of members and no. of non-members visiting the supermarket are almost same.

## 9. No. of females and males visiting the supermarket is almost equal.

## 10. No. of male non-member customers visiting the supermarket is more than the number of female non-member customers visiting the supermarket.

## No. of female member customers visiting the supermarket is more than the number of male member customers visiting the supermarket.

## 11. Fashion accessories are sold more as compared to other product lines.

## Machine learning models:

## 1.Linear regression:

X(dependent variable):Sum of the gross margin percentage of a particular day.

Y(independent variable): Sum of unit price, date(converted to number of days), gender, product type, city, member(Categorical variables converted to continuous).

After cleaning and normalizing the input features using StandardScaler and splitting the data into train and test samples,

the model training is done and we predict the gross margin percentage for the test input.

1. Random Forest Classifier

X(dependent variable):Sum of the gross margin percentage of a particular day.All the float values are converted to integer before applying the model.

Y(independent variable): Sum of unit price, date(converted to number of days), gender, product type, city, member(Categorical variables converted to continuous).

Using Random forest Classifier,we first clean the input data,then split the data into train and test samples ,train the model and predict the gross margin percentage for the test data.

Prediction Accuracy:

**R-squared** (**R2**) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

The r2 score for the Linear Regression model is 0.95 which is really good as the best value for r2 score is 1.Hence the model prediction for gross margin percentage is good.

**Root Mean Square Error** (**RMSE**) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the **regression** line data points are; **RMSE** is a measure of how spread out these residuals are.

The RMSE value for the Linear regression model is 2.91 which is good as the best value for RMSE is 0.

|  |  |
| --- | --- |
| Accuracy(R2 score) | 0.95 |
| Root mean squared error | 2.91 |

Accuracy Score:

In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y\_true.

The Accuracy Score for the Random Forest Model is 39.13% which is pretty good.