SUPERMART SALES ANALYSIS

SUMMATIVE CAPSTONE PROJECT BY MICHELLE ANG

1. Problem statement

2. ER diagram

3. Data Cleaning & Preparation

4. Data Analysis & Dashboard

5. Methodology & Process Workflow

-> Machine Learning

6. Results & 7. Conclusion

01. PROBLEM STATEMENT

SuperMart is interested to understand the demographics of its customers and sales analysis of the products.

Regression Model: Predict the sales amount

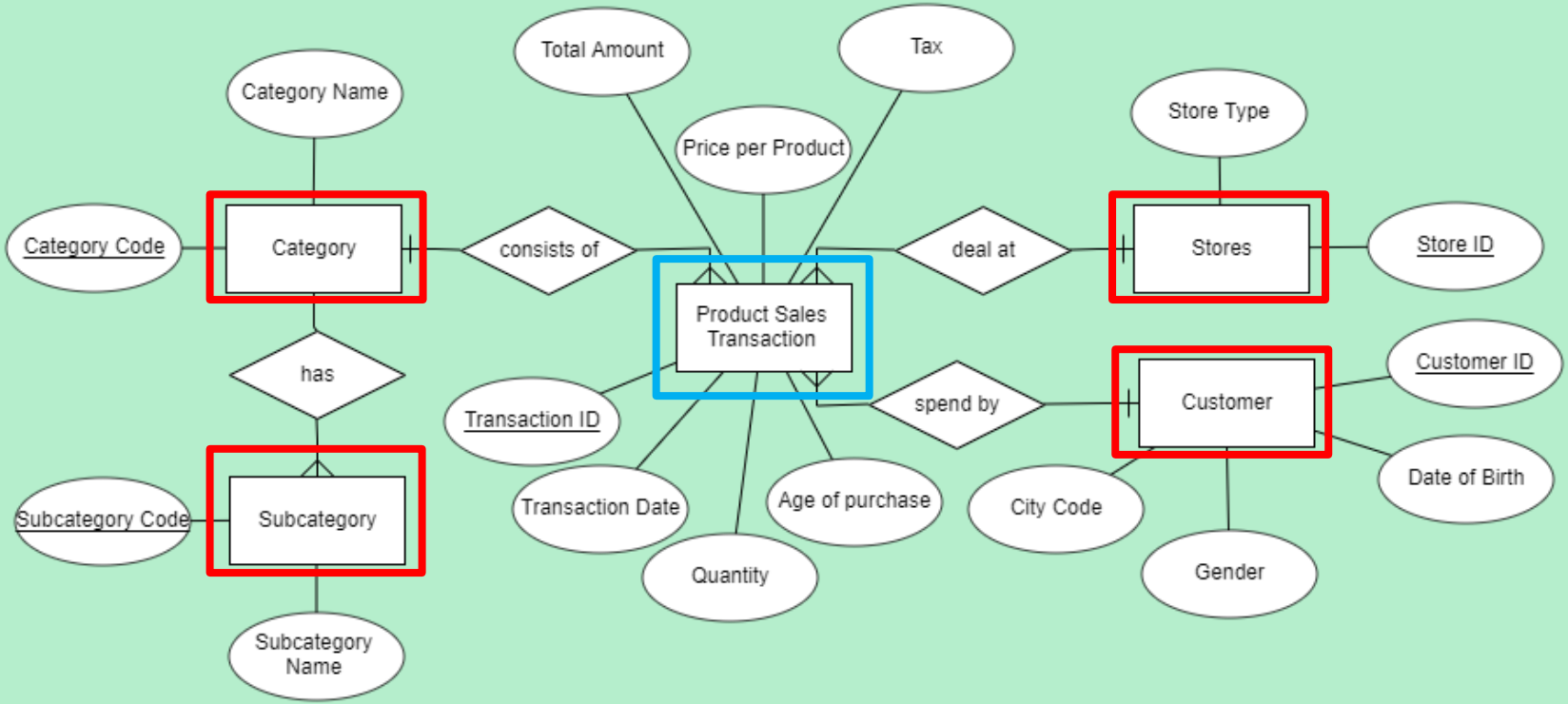
Goal: Able to have profitable business and provide attractive pricing based on its demographics to increase its sales.

Target Audience: Sales Department of SuperMart

02.

ER diagram

ER diagram


Total Amount
Tax
Category Name
Price per Product
Store Type
Category Code
Category
consists of
deal at
Stores
Store ID
Product Sales Transaction
has
Customer ID
spend by
Customer
Transaction ID
Date of Birth
Transaction Date
Age of purchase
City Code
Subcategory Code
Subcategory
Quantity
Gender
Subcategory Name

03.

Data Cleaning & Preparation

Data Cleaning & Preparation

Customer Age at the time of purchases

- Transaction Date minus Birth Date

Remove rows for Qty & Rate

- Ignoring negative data for the analysis: Took out the repeated transactions because of refund order

transaction_id	cust_id	tran_date	prod_cat_code	prod_subcat_code	Qty	Price	Tax	total_amt	Store_type
87125650	268666	09-08-11	4	1	-5	-359.00	188.48	-1983.48	e-Shop
87125650	268666	05-08-11	4	1	5	359.00	188.48	1983.48	e-Shop

Data Cleaning & Preparation

Normalisation- Category & Subcategory

- Create new table for category and subcategory, relabel of their unique identities

Original

prod_cat_code	prod_cat	prod_sub_cat_code	prod_subcat
1	Clothing	4	Mens
1	Clothing	1	Women
1	Clothing	3	Kids
2	Footwear	1	Mens
2	Footwear	3	Women
2	Footwear	4	Kids

After changes

Cat_Code	Category
1	Home and Kitchen
2	Clothing
3	Bags
4	Electronics
5	Books
6	Footwear

Sub_Cat_Code	Sub_Category	Cat_Code
5	Men's Clothing	2
6	Women's Clothing	2
7	Kids' Clothing	2
21	Men's Footwear	6
22	Women's Footwear	6
23	Kids' Footwear	6

Data Cleaning & Preparation

Normalisation- Store

- Using the SQL queries to find out the different distinct store

Adding Primary Key And Foreign Key

- Cust_ID, Cat_Code, Sub_Cat_Code, Store_ID -> Primary Keys
- Add this Primary keys to Product Sales Transaction Table as Foreign Keys

Data Cleaning & Preparation

Import SQL into Python using pyodbc

- Merging Tables for Customer tables and Transaction tables

Removed unneccessary columns for ML

- Unique Customer_ID and Transaction_ID
- Date of Birth and Transaction Date (Year Extracted)
- Price & Tax (colinear to Sales Amount)

04.

Data Analysis & Dashboard

Demographics of SuperMart

Total no. of Transactions

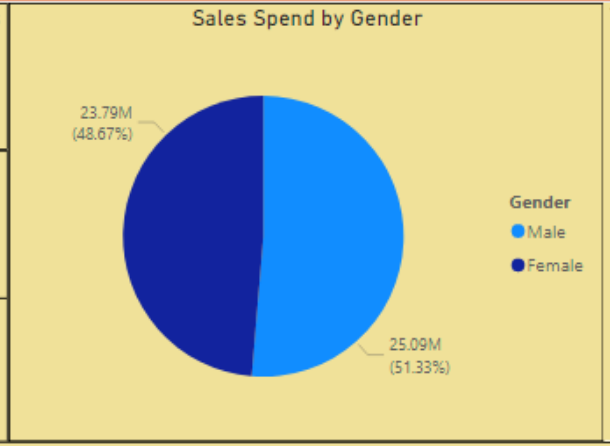
18.82K

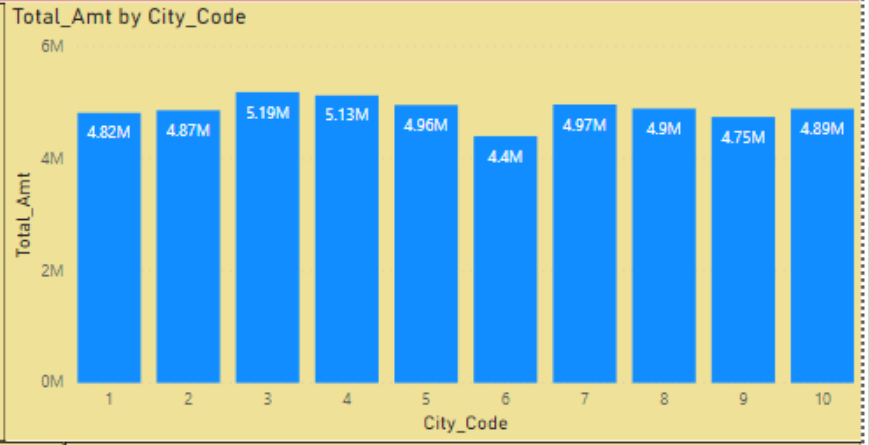
Distinct Customers

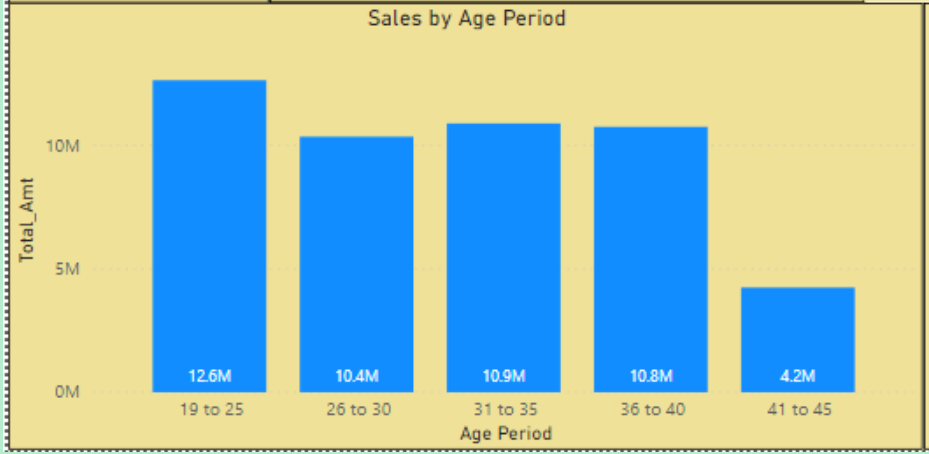
5444

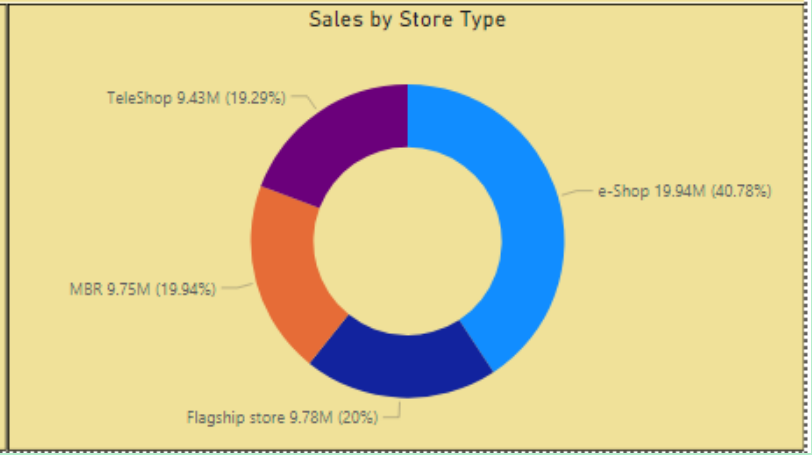
Total Amt

48.90M

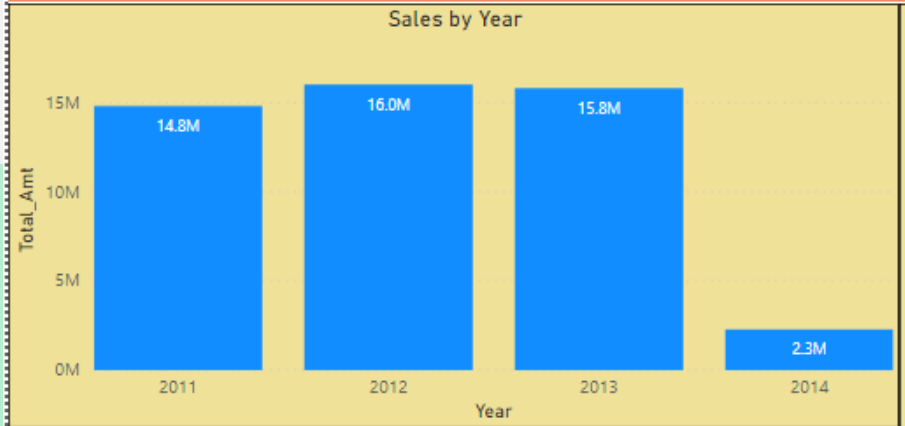


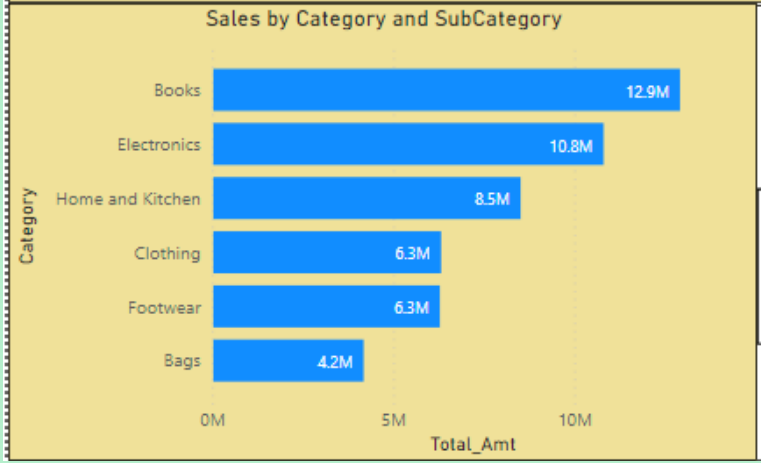


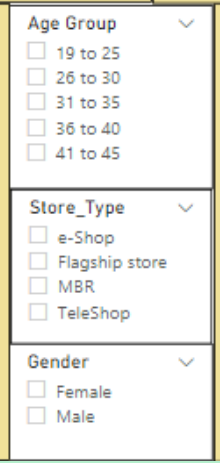


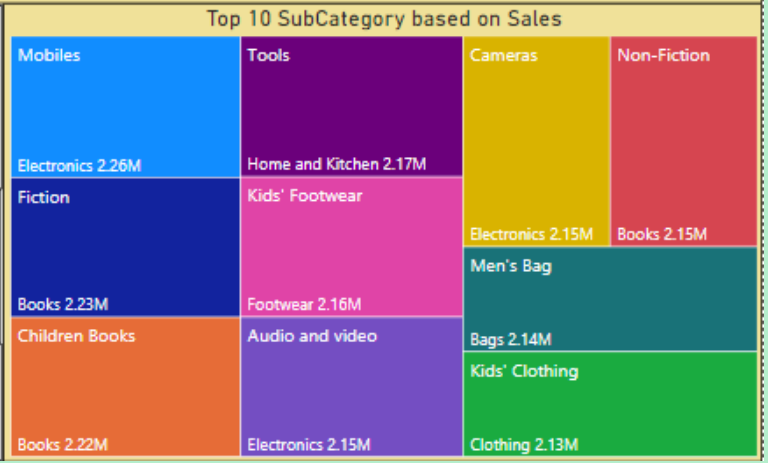


SuperMart's Sales Analysis









05.

Methodology & Process Workflow

Machine Learning

Methodology

Models

Baseline Model: Multiple Linear Regression
Other Models: Random Forest Regressor, Decision Tree Regressor, KNN Regressor, Multi-Level Perceptron Regressor, Gradient Boosting Regressor, Extreme Gradient Boosting

Dataset

Kaggle

Metrics

R2 score, Mean Square Error

Tools

NumPy pandas scikit learn jupyter matplotlib seaborn

Data Transformation

- Nominal Categories Feature (Dummies)

	Gender	Age	Category	SubCategory	Store_Type	Qty	Year
0	M	39	Books	Non-Fiction	e-Shop	2	2013
1	M	39	Clothing	Men's Clothing	e-Shop	1	2013
2	M	38	Clothing	Men's Clothing	TeleShop	3	2012
3	F	21	Books	Fiction	e-Shop	5	2012
4	F	21	Electronics	Mobiles	Flagship store	2	2012

- Split the data into train and test datasets (test size = 0.2)
- Data Normalisation using Standard Scaler
- Finding hyperparameter using Grid Search

06.

Results

Machine Learning

ML Model Training: Multiple Linear Regression

	Model	Training Variance	Testing Variance	Mean Square Error	R2
0	Baseline- Linear Regression	0.864078	0.869686	505011.256143	0.869660

- Result for R2 score is 0.8697 and MSE is 505011

Other ML Models Training

	Model	Training Variance	Testing Variance	Mean Square Error	R2
8	Extreme Gradient Boosting	0.999978	0.999819	699.853837	0.999819
7	Gradient Boosting	0.999908	0.999752	961.296706	0.999752
4	Random Forest	0.999949	0.999641	1390.453929	0.999641
3	Decision Tree	0.999977	0.998542	5651.038443	0.998542

- Extreme Gradient Boosting: Highest R2 and lowest MSE

Extreme Gradient Boosting Model Prediction

My target value is = 833.47

This was the input data:

Category_Bags	0
Category_Books	1
Category_Clothing	0
Category_Electronics	0
Category_Footwear	0
Category_Home and Kitchen	0
SubCategory_Academic	0
SubCategory_Audio and Video	0
SubCategory_Bath	0
SubCategory_Cameras	0
SubCategory_Children Books	0
SubCategory_Comics	0
SubCategory_Computers	0
SubCategory_DIY	0
SubCategory_Fiction	0
SubCategory_Furnishing	0
SubCategory_Kids' Clothing	0
SubCategory_Kids' Footwear	0
SubCategory_Kitchen	0
SubCategory_Men's Bag	0
SubCategory_Men's Clothing	0
SubCategory_Men's Footwear	0
SubCategory_Mobiles	0
SubCategory_Non-Fiction	1
SubCategory_Personal Appliances	0
SubCategory_Tools	0
SubCategory_Women's Bag	0
SubCategory_Women's Clothing	0
SubCategory_Women's Footwear	0
Store_Type_Flagship store	0
Store_Type_MBR	0
Store_Type_TeleShop	0
Store_Type_e-Shop	1
Gender	1
Age	33
Qty	2
Year	2013
Price_*_Age	12672

→ Female

07.

Conclusion

Conclusion

1. E-shop business highest among all store type
2. Age between 19 to 25 years old have a highest purchasing power whereas age between 40 to 45 years old have the lowest purchasing power
3. Mobiles, Fiction books and Children Books are the top 3 subcategory purchases

Extreme Gradient Boosting has the highest R2 (indicates a better fit for the model) and lowest MSE (indicates prediction is closer to actual)

Prediction model in SUPERMART Sales Analysis: Improve business sales as able to predict the pricing for target customers which improved customer satisfaction by having personalised shopping experience/recommendation

Recommendation

1. Will have better analysis during collection process if able to obtain the individual product name, transaction delivery location.
2. Include more features to analyse more advanced factors affecting the sale amounts (ie. Brands, Inflation per Year, Shipping Mode)

THANKS!

https://www.kaggle.com/amark720/retail-shop-case-study-dataset