Regression And Classification On Supermarket Sales Dataset

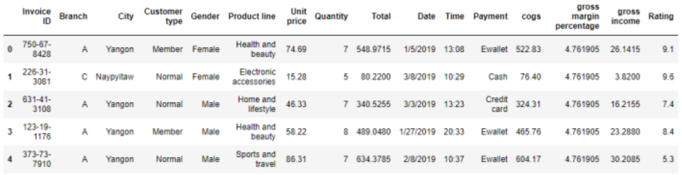
Aaryan Kumar
Department of ECE
Gainesville, USA
aaryan.kumar@ufl.edu

***Abstract*—The purpose of this report is to observe, evaluate and explain different machine learning methodologies used for prediction and classification tasks on the provided supermarket sales data. For the experiment, I have used Multiple linear regression models for regression tasks, and logistic regression, decision tree, and random forest classifiers for the classification tasks. The Multiple Linear Regression models used for regression tasks are of two types, with Lasso regularization and without Lasso regularization. In the case of classification tasks, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier were used. The report also contains evaluation of the said models, and their analysis. This analysis can later be used to increase the profit and efficiency of supermarket operations.**

***Keywords*—machine learning, multiple linear regression, logistic regression, decision tree, random forest.**

I. Introduction

Machine learning provides effective methodologies for regression and classification tasks. In this project I used these methodologies to perform the regression and classification tasks on the supermarket sales dataset. The dataset provided has a total of 1000 samples from three branches of supermarket. There are a total of 16 columns of attributes that were collected. The attributes are of numerical types, and object types. The head of the dataset is shown in Fig. 1



	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3

Figure 1

Different data preprocessing techniques were applied to the dataset as per the requirement of the task, as described later in the report. The task-specific datasets were used, and the machine learning models were trained on the said datasets. The models were saved and tested on the test datasets respectively. A qualitative evaluation of these models was performed, and the observations were recorded.

II. Data Preprocessing

For the data preprocessing task, first an exploratory data analysis was performed. The data was visualized, as shown in Figure 2, and the following was observed-

1. The dataset has a total of 1000 samples. There are a total of 16 attributes, and none of the samples have any missing attribute value.
2. It can also be observed that the dataset has a mix of numerical and categorical attributes, Which, during the data-preprocessing will have to be encoded and transformed to numerical type.
3. Some attributes such as 'Invoice ID', 'gross margin percentage', etc. will have to be dropped due to being irrelevant to the problem.

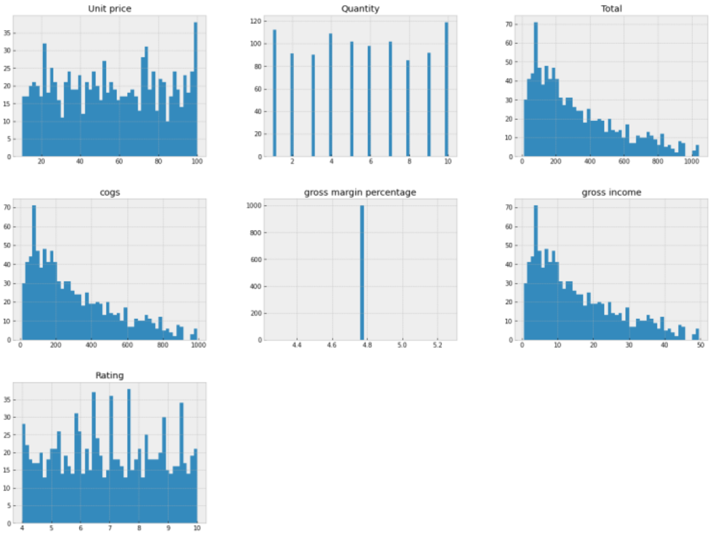


Figure 2

Furthermore, on observing the time and date attributes, I observed that these attributes will have to be encoded as per the requirement.

By looking at the dataset for all the numerical attributes, it is clear that the attributes need to be scaled for an efficient modelling process.

The date attribute was in the mm/dd/yyyy form, which was encoded to day-of-week format using the pandas library. Furthermore, the time attribute was encoded to 4 categories namely, Morning (10:00-12:00), Afternoon (12:01-17:00), Evening (17:01-19:00) and Night (19:01-21:00).

For each specific task, the dataset was preprocessed accordingly, which included dropping unrequired attributes, scaling numerical attributes, encoding categorical attributes, and transforming using pipeline feature of sci-kit library.

These task specific datasets were then saved as csv files, later to be used for training and testing.

III. Machine Learning Model Training

For the project, there are a total of 5 different machine learning tasks, each of which is described below-

A. *Multiple Linear Regression for predicting Gross Income*

The first task was to develop a multiple linear regression model for the predicting Gross income. The dataset created during the data preprocessing step had dropped attributes such as 'Invoice ID', 'Branch', 'City', 'Customer type', 'Gender', 'Total 'Payment', 'cogs', 'gross margin percentage', and 'Rating' for being unimportant for the task. The remaining numerical attributes were scaled, and the categorical attributes were encoded using OneHotEncoder tool of sci-kit library. The gross income was selected as the target attribute.
The model was then trained in two variations – with lasso regularization and without lasso regularization using pipelines.

To find the best parameters, hyperparameter tuning was performed using 5-fold grid-search cross validation. The models with the best estimator was saved as a pickle file, to be later used for testing. The best λ for Lasso regularization was found to be 0.01.

B. *Multiple Linear Regression for predicting Unit Price*

The second task was to develop a multiple linear regression model for predicting Unit price. I used the same dataset used for the prediction of Gross income, with only changing the target attribute to Unit price. The dropped attributes were as 'Invoice ID', 'Branch', 'City', 'Customer type', 'Gender', 'Total 'Payment', 'cogs', 'gross margin percentage', and 'Rating' for being unimportant for the task. In this task also, 2 variations of Multiple linear regression models were trained, viz. with Lasso regularization, and without Lasso regularization. This step was done using pipelines

To find the best parameters, hyperparameter tuning was performed using 5-fold grid-search cross validation. The models with the best estimator were saved as a pickle file, to be later used for testing. The best λ for Lasso regularization was found to be 0.01.

C. *Logistic Regression for classification of gender*

This task is a classification problem. For the data preprocessing step, following attributes were dropped – 'Invoice ID', 'Branch', 'City', 'Customer type', 'Unit price', 'Quantity', 'Total', 'Date', 'Time', 'cogs', 'gross margin percentage', 'Rating' for being unimportant for the task. The numerical attributes were scaled, and the categorical attributes were encoded. The gender was set as the target attribute, having 2 classes.
The model pipeline was then created which had 2 steps, viz. Polynomial feature Extraction, and Logistic regression. Polynomial feature extraction was set with a degree of 2 to extract polynomial features of the dataset and feed it to logistic regression.
The trained model was then saved as a pickle file.

D. *Logistic Regression for classification of customer type*

This task is a classification problem. For the data preprocessing step, following attributes were dropped - 'Invoice ID', 'Branch', 'City', 'Unit price', 'Quantity', 'Total', 'cogs', 'gross margin percentage', 'Rating', 'Product line', 'gross income', 'Payment' for being unimportant for the task. The numerical attributes were scaled, and the categorical attributes were encoded. The customer type attribute was set as the target attribute, having 2 classes.
The model pipeline structure for this task was similar to the previous gender classification structure. It had 2 steps, viz. Polynomial feature Extraction, and Logistic regression. Polynomial feature extraction was set with a degree of 2 to extract polynomial features of the dataset and feed it to logistic regression.
The trained model was then saved as a pickle file.

E. *Predicting Day of Purchase*

For the task of predicting day of purchase (essentially a classification task), 2 different classifiers were trained, viz. Decision Tree Classifier, and Random Forest Classifier.

i. Decision Tree Classifier

The decision tree classifier was built using a pipeline.5-fold Grid-search cross validation method was used to find the best parameters for the model. The scoring method used for the grid search cross validation was accuracy because of the task being a classification task. The model with the best estimator was saved as a pickle file.

ii. Random Forest Classifier

The random forest classifier was built using a pipeline. using a pipeline. 5-fold Grid-search cross validation method was used to find the best parameters for the model. The scoring method used for the grid search cross validation was accuracy because of the task being a classification task. The model with the best estimator was saved as a pickle file.

IV. Machine Learning Model Testing Results And Evaluation

All the saved machine learning models were tested and evaluated and following was observed

A. *Multiple Linear Regression for predicting Gross Income*

The saved models for multiple linear regression for Gross income were loaded and used to predict target values for the test data. To evaluate the models, coefficient of determination, r^2 was calculated, and its 95% confidence interval was calculated. Furthermore, the effect of attributes on gross income was also observed. The observations are as follows-

1. Multiple linear regression with lasso regularization-
   λ= 0.01
   r^2 score = 0.90332
   95% confidence interval = (-0.48747, 7.26488)

2. Multiple linear regression without lasso regularization-
   r^2 score = 0.9048
   95% confidence interval = (-0.38164, 7.72420)

Furthermore, I observed from the correlation matrix that Gross income has highest positive correlation with *Quantity*, followed by *Unit price, Electronic accessories and Thursday*.

B. *Multiple Linear Regression for predicting Unit Price*

The saved models for multiple linear regression for Unit Price were loaded and used to predict target values for the test data. To evaluate the models, coefficient of determination, r^2 was calculated, and its 95% confidence interval was calculated. Furthermore, the effect of attributes on gross income was also observed. The observations are as follows-

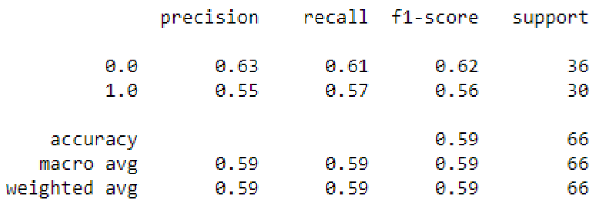
3. Multiple linear regression with lasso regularization-
   λ= 0.01
   r^2 score = 0.84176
   95% confidence interval = (-1.30533, 13.620150)

4. Multiple linear regression without lasso regularization-
   r^2 score = 0.79427
   95% confidence interval = (-0.69960, 1.62586)

Furthermore, I observed from the correlation matrix that Unit price has highest positive correlation with *gross income, followed by electronic accessories, Food and beverage, Quantity and Sunday.*

C. *Logistic Regression for classification of gender*

The saved models for logistic regression for gender classification were loaded and used to predict target values for the test data. A classification report was generated, as shown in Figure 3. Furthermore, the parameter values for the attributes were plotted (as shown in Figure 4), and most informative attributes were found.



```
              precision    recall  f1-score   support

         0.0       0.63      0.61      0.62        36
         1.0       0.55      0.57      0.56        30

    accuracy                           0.59        66
   macro avg       0.59      0.59      0.59        66
weighted avg       0.59      0.59      0.59        66
```

Figure 3

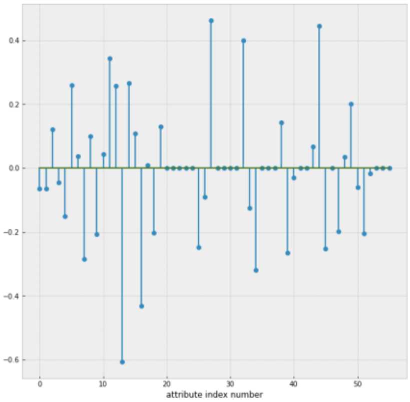


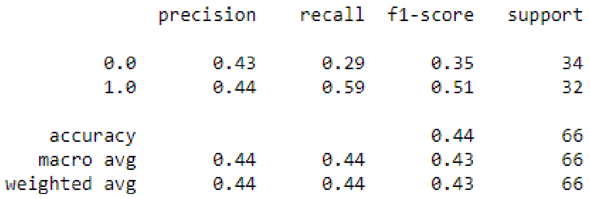
Figure 4

By observing the absolute values of the coefficients of the attributes (and their 2nd order interaction), I found that following attributes were most informative-

1. gross income + Food and beverages
2. Electronic accessories + Ewallet
3. Health and beauty + Credit card
4. gross income + Sports and travel
5. Fashion accessories + Cash

D. *Logistic Regression for classification of customer type*

The saved models for logistic regression for customer type classification were loaded and used to predict target values for the test data. A classification report was generated, as shown in Figure 5. Furthermore, the parameter values for the attributes were plotted (as shown in Figure 6), and most informative attributes were found.



```
              precision    recall  f1-score   support

         0.0       0.43      0.29      0.35        34
         1.0       0.44      0.59      0.51        32

    accuracy                           0.44        66
   macro avg       0.44      0.44      0.43        66
weighted avg       0.44      0.44      0.43        66
```

Figure 5

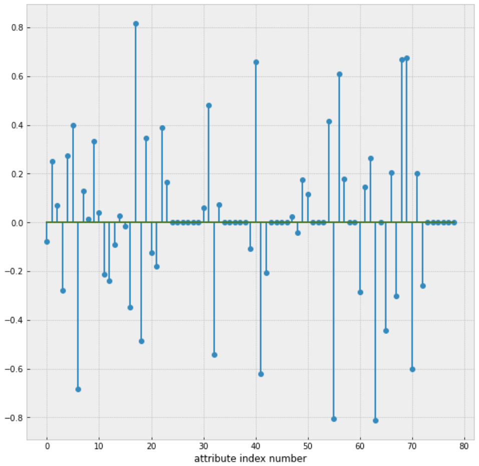


Figure 6

By observing the absolute values of the coefficients of the attributes (and their 2nd order interaction), I found that following attributes were most informative-

1. gender + Monday
2. Monday + Sunday
3. night + Friday
4. Monday
5. Wednesday + Thursday

E. Predicting Day of Purchase

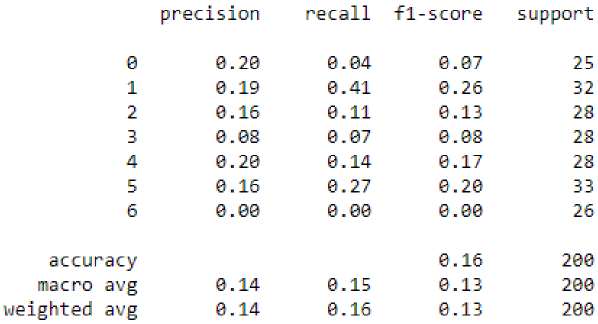
The saved decision tree classifier model and random forest classifier model were loaded and tested on the test data. The classification reports for both classifiers were generated. The accuracy was used as the success metric, and its 95% confidence interval was calculated. Following are the evaluations of both classifiers-

1. Decision Tree Classifier

   Accuracy = 16%

   95% Confidence interval = (0.11357, 0.22642)

   Classification report (Figure 7)



```
              precision    recall  f1-score   support

           0       0.20      0.04      0.07        25
           1       0.19      0.41      0.26        32
           2       0.16      0.11      0.13        28
           3       0.08      0.07      0.08        28
           4       0.20      0.14      0.17        28
           5       0.16      0.27      0.20        33
           6       0.00      0.00      0.00        26

    accuracy                           0.16       200
   macro avg       0.14      0.15      0.13       200
weighted avg       0.14      0.16      0.13       200
```

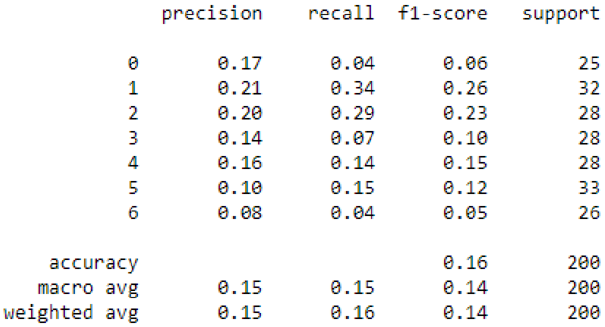
Figure 7

2. Random Forest Classifier

   Accuracy = 16%

   95% Confidence interval = (0.12305, 0.196940)

   Classification report (Figure 8)



```
              precision    recall  f1-score   support

           0       0.17      0.04      0.06        25
           1       0.21      0.34      0.26        32
           2       0.20      0.29      0.23        28
           3       0.14      0.07      0.10        28
           4       0.16      0.14      0.15        28
           5       0.10      0.15      0.12        33
           6       0.08      0.04      0.05        26

    accuracy                           0.16       200
   macro avg       0.15      0.15      0.14       200
weighted avg       0.15      0.16      0.14       200
```

Figure 8

Conclus*ion*

The different machine learning models were trained, tested, and evaluated on the supermarket sales data for the tasks of regression, and classification. For Multiple linear regression models, lasso regularization gave a better result than their counterparts. For logistic regression models, I was able to observe and plot parameters values for all the attributes (including their 2nd order interaction). Furthermore, while testing for the same classification task, random forest classifier performed better than the decision tree classifier. In this project I was able to implement and evaluate different machine learning models and methodologies.

Acknowledgment

I would like to acknowledge my professor, Dr. Catia Silvia, and course TA Jackson Cornell for helping and providing valuable insight for completion of this project.

References

[1] EEL 5934 Course Lectures, office hours and Course notes

[2] EEL 5934 discussions on slack, and during lectures/office hours

[3] Scikit-learn documentation (https://scikit-learn.org/stable/index.html)