Building & Evaluating ML Algorithms

Note: Project 1

Xiangsheng Chai
ECE Department
UFID: 44036693

***Abstract*—In this project, I will use a supermarket sales dataset to build machine learning model and evaluate the model's performance. First, I will use kinds of preprocessing methods to deal with the original data. Then, I apply machine learning model to make regression and classification on the data set. For regression tasks, I apply linear regression model with and without the Lasso regularization. For classification tasks, I apply logistic regression, decision tree and random forest algorithm. After training, evaluating metrics are applied to estimate the performance of the model and seeing the relationship among every features.**

Keywords—Supermarket sales dataset, Regression, Classification, Performance, Relationship.

I. INTRODUCTION

Everyone needs to go to the supermarket every week. The growth of supermarkets in most populated cities are increasing and market competitions are also high. In this project, our goal is to use machine learning algorithms to do regression and classification tasks, and use the performances of each model to analyze the relationships among every attribute of the dataset. In this project, we need to learning how to do the data preprocessing and data cleaning, how to select the best model, how to implement performance evaluation metrics and evaluate results and how to report observations, propose business-centric solutions and propose mitigating strategies.

The dataset we use in this project is 'supermarket sales' dataset. It is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data. The dataset contains totally 16 attributes, and the attribute description is as following:

1. **Invoice id**: Computer generated sales slip invoice identification number.
2. **Branch**: Branch of supercenter (3 branches are available identified by A, B and C).
3. **City**: Location of supercenters.
4. **Customer type**: Type of customers, recorded by Member for customers using member card and Normal for without member card.
5. **Gender**: Gender type of customer.
6. **Product line**: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel.
7. **Unit price**: Price of each product in US dollars.
8. **Quantity**: Number of products purchased by customer.
9. **Total**: Total price including tax.
10. **Date**: Date of purchase (record available from January 2019 to March 2019).
11. **Time**: Purchase time (10am to 9pm).
12. **Payment**: Payment used by customer for purchase (3 methods are available - Cash, Credit card and Ewallet).
13. **COGS**: Cost of goods sold.
14. **Gross margin percentage**: Gross margin percentage.
15. **Gross income**: supercenter gross income in US dollars.
16. **Rating**: Customer stratification rating on their overall shopping experience (on a scale of 1 to 10).

There are totally 6 questions for us to solve. The questions are listed below:

1. Apply the necessary data preprocessing using scikit-learn pipelines. Justify all choices.
2. Train a multiple linear regression with and without Lasso regularization to predict gross income. And analyze the performance with coefficient of determination r^2 and its 95% confidence interval.
3. Train a multiple linear regression with and without Lasso regularization to predict Unit price. And analyze the performance with coefficient of determination r^2 and its 95% confidence interval.
4. Train a logistic regression to classify gender and study the relationship between attributes. Analyze the

parameters values by visualization techniques and select the most informative attribute.

5. Train a logistic regression to classify customer type and study the relationship between attributes. Analyze the parameters values by visualization techniques and select the most informative attribute.
6. Train a classifier to predict the day of purchase. And analyze the performance with accuracy and its 95% confidence interval.

II. Implementations

A. Preprocessing

For this project, there are 2 useless attributes (Invoice id and gross margin percentage), 6 numerical attributes and 8 categorial attributes. Because the dataset contains no none elements, the dataset does not need any cleaning steps. For numerical attributes, standard scaler is applied. For categorial attributes, ordinal encoder is applied on attribute 'Gender' and 'Customer type', and one hot encoder is applied for the others.

B. Regression tasks

In regression tasks, our goal is to use linear regression with and without Lasso regularization to predict gross income and unit price.

The objective function of linear regression is :

$$J(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - f(\phi(x_n), \mathbf{w}))^2 = \frac{1}{2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|_2^2$$

And with Lasso regularization, it will become:

$$J(\mathbf{w}) = \|\mathbf{t} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

The difference between linear regression with and without lasso regularization is that lasso regularization can restrict some of the parameters to zero. It means lasso regularization can exclude some features which are less related by setting their parameters to 0.

Linear regression algorithm and lasso regression algorithm from scikit-learn are used to build the model and coefficient of determent will be use for evaluating.

C. Classification tasks

In classification tasks, logistic regression is applied in question 4 and 5, and in question 6, logistic regression, decision tree and random forest are chosen to classify the dates.

The objective function of logistic regression is:

$$J(\mathbf{w}, w_0) = \sum_{i=1}^{N} -t_i \log \phi(z_i) - (1 - t_i)\log(1 - \phi(z_i))$$

It is also known as cross entropy, and it is the negative value of likelihood to the data. We need to maximize the likelihood, that means minimize the objective function. In order to do that, gradient descent will be applied.

In the question 4 and 5, we combine the logistic regression and polynomial features and analyze the parameters values to see the contribution of each attribute. In question 6, logistic regression with ridge regularization is applied.

Decision tree uses the information gain to measure the quality of the split. It is a non-parametric discriminative classifier. Hence, we do not need to analyze any parameters on it.

Random forest is the application of bagging, which uses decision tree as the estimator to make an ensemble.

Logistic regression, decision tree and random forest algorithm from scikit-learn are used to build the model. For question 4 and 5, we plot the parameters values to visualize the contribution of each attribute. For question 6, accuracy score, 95% confidence interval and confusion matrix are applied to determine the best model of classification.

III. Experiment

A. Problem 1

The supermarket sales dataset totally contains 16 attributes. Firstly, I dropped attributes "Invoice id" and "Gross margin percentage". This is because data in "Invoice id" are just a string of useless numbers, and data in "Gross margin percentage" are all the same. They cannot make contribution for training model.

Then, library pandas was used to transform "Date" to the respective day of the week and "Time" to morning, afternoon, evening and night. After that, we split the data into training set and test set

Finally, standard scaler was applied for numerical attributes, ordinal encoder for categorial attributes which only contain two classes, and one hot encoder for the other categorial attribute. All these transforming steps are combined in one pipeline. After fitting the original dataset, we get our prepared dataset, and it is ready for training.

B. Problem 2

In problem 2, attributes "Unit price", "Quantity", "Date", "Time" and "Product line" are selected as input features. The output is the prediction of "Gross income".

There are two regression algorithms, linear regression and lasso regression. For lasso regression, two strategies of hyperparameters tuning are applied in order to find the best value of regularization parameter λ.

TABLE I. BEST λ IN DIFFERENT CV STRATEGY

CV Strategy	λ
Grid Search CV	0.00498
Random Search CV	0.00115

Coefficient of determination r^2 and its 95% confidence interval are reported as the evaluating metrics. The evaluating result of training set is as following:

TABLE II. EVALUATION IN TRAINING SET

MODEL	r^2	95% CI
LINEAR	0.883	[0, 1]
LASSO WITH GRIDSEARCHCV	0.882	[0, 1]
LASSO WITH RANDOMSEARCHCV	0.883	[0, 1]

The evaluating result of test set is as following:

TABLE III. EVALUATION IN TEST SET

MODEL	r^2	95% CI
LINEAR	0.918	[0, 1]
LASSO WITH GRIDSEARCHCV	0.918	[0, 1]
LASSO WITH RANDOMSEARCHCV	0.918	[0, 1]

From the evaluation results, each model has a great coefficient of determination. Compared to others, lasso regression with grid search CV strategy has the most compact confidence interval. In this case, lasso regression with grid search CV was chosen as the best model and $\lambda = 0.00498$ was chosen as the best values of hyperparameter.

By analyzing the parameters' values of the best model, only "Unit price", "Quantity", "Tuesday" and "Wednesday" in "Date" and "Afternoon" in "Time" are contributing for regression, other attributes are excluded. Among these useful attributes, "Unit price" and "Quantity" are most informative in predict "Gross income" because of their greater parameters' values.

C. *Problem 3*

Problem 3 used the same strategy of training, but the output is changed to "Unit price" and the input is substituted to "Gross income". The necessary tables are shown below:

TABLE IV. BEST λ IN DIFFERENT CV STRATEGY

CV Strategy	λ
Grid Search CV	0.00618
Random Search CV	0.00115

TABLE V. EVALUATION IN TRAINING SET

MODEL	r^2	95% CI
LINEAR	0.769	[0, 1]
LASSO WITH GRIDSEARCHCV	0.768	[0.710, 0.892]
LASSO WITH RANDOMSEARCHCV	0.769	[0.039, 1]

TABLE VI. EVALUATION IN TEST SET

MODEL	r^2	95% CI
LINEAR	0.831	[0.818, 1]
LASSO WITH GRIDSEARCHCV	0.829	[0.804, 1]
LASSO WITH RANDOMSEARCHCV	0.831	[0.820, 1]

From the tables above, $\lambda = 0.00618$ and lasso regression with grid search CV are chosen for the best model. This is because, grid search CV results has the most compact confidential interval.

By analyzing the parameters' values of the best model, only "Gross income", "Quantity", "Tuesday" and "Wednesday" in "Date" and "Afternoon" in "Time" are contributing for regression, other attributes are excluded. Among these useful attributes, "Gross income" and "Quantity" are most informative in predict "Unit price" because of their greater parameters' values.

D. *Problem 4*

In problem 4, I use logistic regression with polynomial feature in degree 2 as my classifier. After transformation by pipeline I set, I got totally 56 combinations of attributes. The confusion matrix of test set is shown below:

TABLE VII. CONFUSION MATRIX OF TEST SET

Gender	Male	Female
Male	21	9
Female	20	8

And the parameters values figure is :

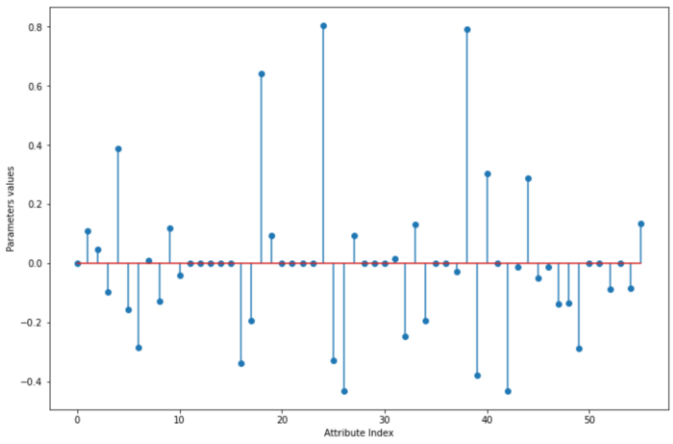

Fig. 1. Parameters values for all attributes on "Gender" classifier

From the figure, we can see some parameters values are equal to 0, which means these attributes are excluded. Most of these excluded attributes are the inter combinations of "Date"

and “Time”, like “ Monday + Tuesday” and “Morning + Night”. Among these values, attribute “ Fashion accessories + Cash” that has greatest parameter value in absolute value is the most informative attribute.

E. Problem 5

In problem 4, I use logistic regression with polynomial feature in degree 2 as my classifier. After transformation by pipeline I set, I got totally 79 combinations of attributes. The confusion matrix of test set is shown below:

TABLE VIII. CONFUSION MATRIX OF TEST SET

Customer type	Normal	Member
Normal	10	22
Member	15	11

And the parameters values figure is :

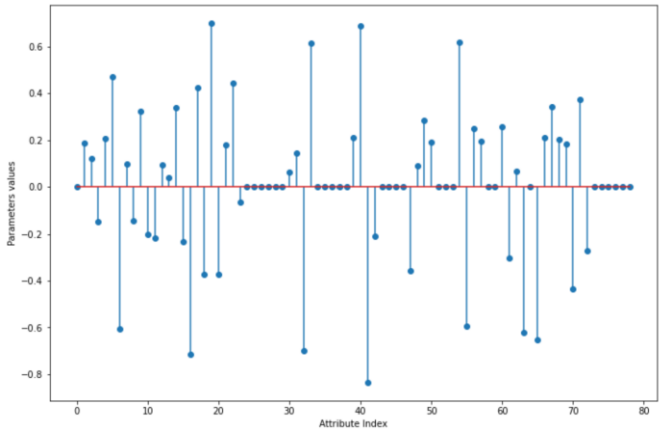

Fig. 2. Parameters values for all attributes on “Customer type” classifier

The same as the problem 4, most of the inter combinations of “Date” and “Time” are excluded. And the most informative attributes in this model is “ Tuesday + Evening”.

F. Problem 6

In problem 6, logistic regression, decision tree and random forest were applied for classifying. Then, grid search cross validation was used for hyperparameters tuning.

For logistic regression, the best hyperparameter is C=0.7. For random forest, the best number of estimators is 200. And for decision tree, the best hyperparameters are shown below:

TABLE IX. HYPERPARAMETS OF DECISION TREE

Criterion	Entropy
Max Depth	11
Min samples leaf	5
Min sample split	3

Accuracy, confusion matrix and 95% confidence interval of accuracy were applied for evaluating the performance and determining the best model of classification. The results of evaluating are shown below:

TABLE X. PERFORMANCE OF EACH MODEL

	Logistic Regression	*Decision Tree*	*Random Forest*
Accuracy in train	0.236	0.4175	1.0
Accuracy in test	0.115	0.15	0.14
Confidence interval in train	[0.136, 0.157]	[0.150, 0.202]	[0.120, 0.155]
Confidence interval in test	[0.055, 0.195]	[0.062, 0.198]	[0.059, 0.131]

From the Table X, we can see all of the models have a bad accuracy on test set and random forest model is overfitted. It means, every model cannot precisely classify the date of purchase. It could be the less of the data or the incapacity of the model. Compared to logistic regression, decision tree has a better accuracy on test set and a more compact confidence interval. Compared to random forest, decision tree is not overfitting. In this case, even though three models are all bad in performance, decision tree is better than the other two models.

IV. CONCLUNSION

In regression task, we can see lasso regularization can effectively exclude the useless attributes in order to make prediction more accurate. And compared to random search, grid search could take more time but find the better hyperparameters values.

In classification task, the performance of logistic regression model, decision tree model and random forest model are not very well. It seems to be the capacities of three models are not powerful enough for this supermarket sales dataset, we need more complicated model.