THE REPORT OF FLIP00 FINAL PRESENTATION

ZHAOYANG WANG

Abstract. This report contains five parts.The first part will introduce the problem, describe the data and analyzes the problem. And the Second will do the statistic of the data, visualize the data to find some potential relationships between the attribute values. Third, this part will explain the method that i use.The Fourth will introduce the experiment and analysis of the algorithm and result.The last one is conclusion.

Contents

Date: 2019-11-27.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

1.1. **Problem Statement.**

You are given 5 years of store-item sales data, and asked to predict 3 months of sales for 50 different items at 10 different stores.

1.2. **Data List.**

The data contains date of the item sold at a store ,And the name of the store,And the name of the item, and the sales of every store on a particular day.

date: - Date of the sale data. There are no holiday effects or store closures.
store: - Store ID.
item: - Item ID.
sales: - Number of items sold at a particular store on a particular date.

1.3. **Problem Analysis.**

This is a question that asks us to predict sales for three months based on sales that have been available for three years. And We can analyze that this is a regression problem.

2. Exploratory Data Analysis

2.1. **Data Information.**

From the table 1, we can clearly understand the situation of the data set. And we can see the sales volume and sales time of each product in each store during the term of 2013-1-1 to 2017-12-31.

Table 1. Data Information

	date	store	item	sales
0	2013-01-01	1	1	13
1	2013-01-02	1	1	11
2	2013-01-03	1	1	14
3	2013-01-04	1	1	13
4	2013-01-05	1	1	10

2.2. **Data Visualization.**

By using the matplotlib to plot the photoes which describe the sale pattern. For example, The distribution of sales volume ,total sales of the stores, total sales of the items, store's performance over the time, all items' performance over the time ect.

From the figures 1 we can know that 2nd store is the topper and 7th store is the least revenue generating one; From figure 2 we can know that more and more sales volume is belong to [0,50]; From figure 3 we can know that 2nd store is the topper of the all stores; From figure 4 we can know the total sales of all items; From figure 5 we can know every stores sales changing overtime; From figure 6 we can know every item sales changing overtime; From figure 7 and 8 we can know individual pattern of item's sale and store's sale.

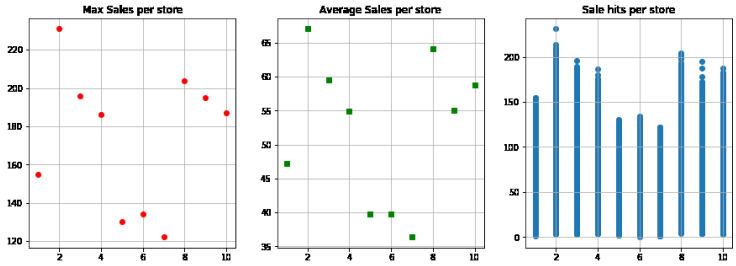

FIGURE 1. Displaying the sale pattern

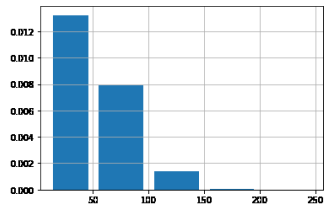

FIGURE 2. sales volume's distribution

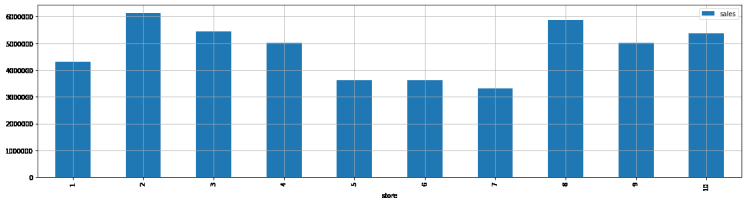

FIGURE 3. The total sales of all stores

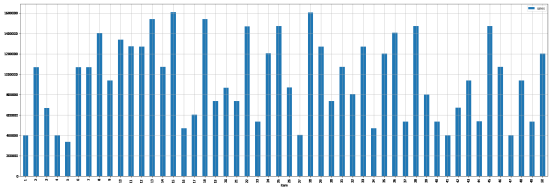

FIGURE 4. The total sales of all items

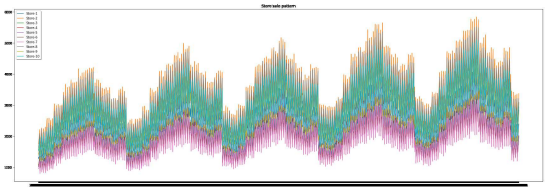

FIGURE 5. The performance of all stores

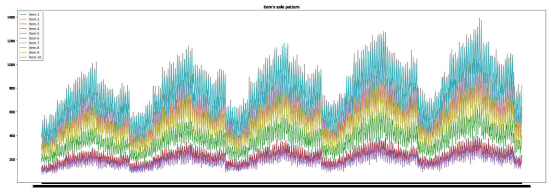

FIGURE 6. The performance of all stores items

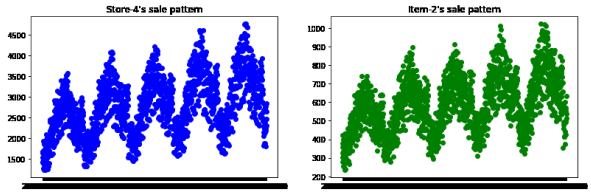

FIGURE 7. The performance of the individual store and item

2.3. **Data Preparation.**

Through the data visualization before, we can intuitively recognize the changes in sales. However, to forecast sales for the next three months, we need to extract some new features. From the previous figure we can see that the sales are related to the characteristics of the year, month, season, etc., so we can add some new features.

The new features that are specifically added are as follows:

dayofweek: - Indicates that this day is the day of the week. Monday is indicated by 0, Tuesday is indicated by 1, and so on. Sunday is indicated by 6.

is_weekend: - Determine if this day is a weekend. The weekend is indicated by 1 and the working day is represented by 0.

day: - Judging this day is the first few days of this month.

year: - Judging year.

dayofyear: - Judging that day is the first few days of the year.

weekofyear: - Judging that the day is the first few weeks of the year.

sales_mean_lag_90: - Calculate 90 days from the day before, and then start from this day, the average of the first seven days.

sales_std_lag_90: - Calculate 90 days from the day, and then start from this day, the standard deviation of the first seven days.

3. METHODS

There are many machine learning methods for solving regression problems. This moment I will chose the lightGBM model.

3.1. **Trainning model.**

The lightGBM is light Gradient Boosting Machine which is a gradient boosting framework that uses tree based learning algorithms.

It is designed to be distributed and efficient with the following advantages:

: Faster training speed and higher efficiency.

: Lower memory usage.
: Better accuracy.
: Support of parallel and GPU learning.
: Capable of handling large-scale data

Select all the features except for the target variable. And Train LightGBM model.

3.2. **Evaluate model.**

Verify the training results using the smape model. Smape is Symmetric Mean Absolute Percentage Error.

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

4. Experiment and Analysis

First i divide training data and test data by using sklearn, and then i do a model training. The next step is model prediction, and then i do a model evaluation by using SMAPE. Third, re-train the model and the Feature importances will be get. Last ,I will make test predations

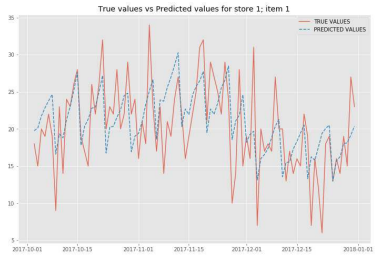

Figure 8. The value vs predict

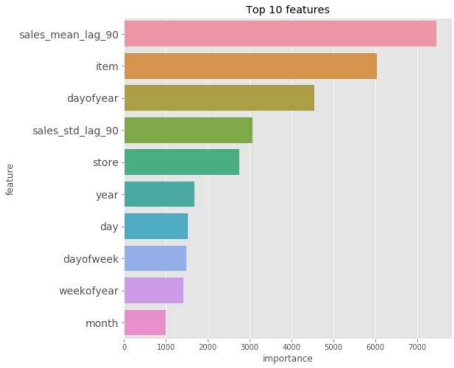

Figure 9. Top 10 importabce features

ggtgtyfytfyt

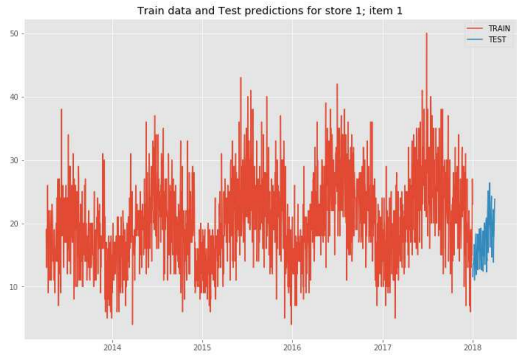

FIGURE 10. The test prediction

Figure 9 is the first prediction based on the model. Figure 10 shows ten features with high feature importance. Figure 11 is a new prediction based on the model to get the results needed for this problem.

5. CONCLUSION

Data Analysis Using visual methods to find connections and features within the dataset.
Feature Engineering Find and extract important features.
Modeling Choose the suitable model parameters.
Prospcting I woule like to select multiple models for comparison later.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy