ITP4514 Artificial Intelligence and Machine Learning

Group Assignment

**[Inventory Demand Forecasting]**

|  |  |  |
| --- | --- | --- |
| **Group Member** | **Student Name** | **Student Id.** |
| 1 | Lo Ching Wa | 200175988 |
| 2 | Li Chi Kit | 200051863 |
| 3 | Luk Chin Pang | 200013999 |
| 4 | Ho Chun Man | 200227387 |

Index

Index 2

Introduction / Background 3

Problem Formulation 4

Methodology 5

Findings & Results 7

Summary 14

References 15

**Appendix** 16

# Introduction / Background

*Definition*

Inventory demand forecasting is the process of predicting customer demand for an inventory item over a defined period of time. Accurate inventory demand forecasting enables a company to hold the right amount of stock, without over or under-stocking, for optimum inventory control. [1]  
It uses factors such as sales history and trends, average lead time, demand, reorder point, and safety stock to predict inventory levels. [2]

To study inventory demand forecasting, a study case is implemented in our report to help us visualize this concept.

*Study Case [3]*

An Indian retail chain has decided to manage the inventory across 3 states in India: Maharashtra, Telangana, and Kerala. In each store, they stock products across various categories, such as fast moving consumer goods, eatables / perishables, food and drinks, and others. The company would like to meet the demand to not lose potential revenue.

Files:

* Train dataset: included 26 months of sales record for training
* Test dataset: 1 month to be predicted

*Outline*In the next section, we would discuss the importance of demand forecasting. Next, our methodology would be listed on Section 3. Detailed findings and results would be discussed on Section 4. Lastly, we would conclude in the last section.

# Problem Formulation

**Why Inventory demand forecasting is important.**

This task is fundamental, crucially important to running a business smoothly and making sound operational decisions, and notoriously difficult to perform accurately. Demand forecasting is so pivotal because it allows a business to set correct inventory levels, price their products correctly, and understand how to expand or contract their future operations. Poor forecasting can lead to lost sales, depleted inventory, unhappy customers, and millions in lost revenue. [4]

**Common demand forecasting problems**

* Forecasting demand too low

Planning to meet demand that’s lower than what actually needs to be met will leave you scrambling to catch up. When something unexpected happens and you don’t have enough materials and labor, you end up shipping orders late because you have to scramble to find workers and the things you need to make the products.

* Forecasting demand too high

Planning to meet demand that is higher than what actually materializes will result in overstaffing and excess inventory. This leads to overtime costs and potentially storage costs for the extra materials you have on hand.

* Forecasting demand too early

Forecasting your demand too early means you’ll almost certainly miss the mark. The further out you plan, the harder it is to be precise with your predictions.

* Relying too rigidly on forecast demand

Having a plan is great but sticking to it so rigidly that you can’t respond to the reality around you is a surefire way to get your business into trouble. Rigidity costs you time, money, and opportunities. But, depending too much on forecast demand can make it so that companies won’t look at the situation unfolding around them and make a change to respond to the situation. [5]

# Methodology

**There are a few methods help us to predict the Inventory demand forecasting result. We found some dataset to predict a continuous value to find next period result of Inventory demand forecasting.**

1. **Simple Linear Regression**

The statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Simple Linear Regression algorithm assume that two variables are linearly related.

**Linear regression is called simple if only work with one independent variable (x).**

**Formula : f(x) = mx+b**

**Cost Function**

We can measure the accuracy of our linear regression algorithm using the mean squared error cost function.

Mean squared error measures the average squared distance between the predicted output and the actual output.

**It is commonly used as a quantitative way to determine the underlying trend and when prices are overextended.**

1. **Confusion Matrix**

It is an evaluation of classification model.

**Dataset Description**

train\_data.csv

* date : The date for which the observation was recorded
* product\_identifier : The id for a product
* department\_identifier : The id for a specific department in store
* category\_of\_product : The category to which a product belongs
* outlet : The id for a store
* state : The name of the state
* sales : The number of sales for the product

product\_prices.csv

The prices of products at each store for each week

date\_to\_week\_id\_map.csv

The mapping from a date to the weekID

sample\_submission.csv

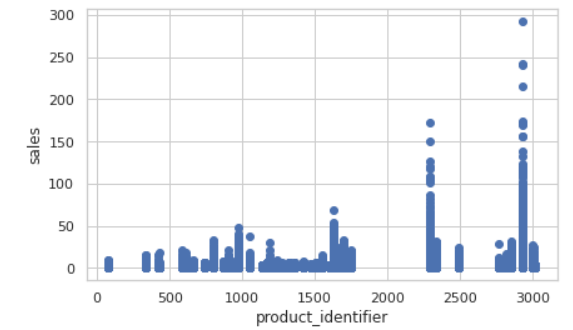
The format for submissions

test\_data.csv

This file has all the attributes of the train\_data.csv file excluding the sales target column

# Findings & Results

To begin with, we found that the linear regression model is suitable for our topic to predict the inventory sold. Fortunately, we found a vast and useful data set from [www.kaggle.com](http://www.kaggle.com), but the most important that we spend a lot of time finding some documents for researching how can be able to understand the source code to build up a linear regression model.

* **Python matplotlib.pyplot.scatter()**

Firstly, we understood a scatter plot is a diagram where each value in the data set is represented by a dot. For example, plt.scatter(dataset[column] , dataset['sales']) this line of source code the dataset[column] represent the x-axis (product\_identifier) and dataset['sales'] represent the y-axis (sales), so it can be seen that the relation between product\_identifier and sales, and here is explaining what products have the high amount of sales in a month.

* **Python | Pandas dataframe.corr()**

一張含有 文字 的圖片

自動產生的描述

一張含有 桌 的圖片

自動產生的描述

Secondly, we understood the “corr()” is meaning correlation, it is used to find the pairwise correlation of all columns in the dataframe. Any na values are automatically excluded. For any non-numeric data type columns in the dataframe it is ignored. Also, we knew the DataFrame.corr() has 3 methods to calculate the correlation coefficient between columns such as pearson for calculating the standard correlation coefficient, kendall for calculating the Kendall Tau correlation coefficient and spearman for calculating the Spearman rank correlation.

In our case, we use the method of pearson corr() function to find the correlation among the columns, here have six numeric columns product\_identifier, department\_dientifier, outlet, sales, week\_id, sell\_price in the dataframe, respectively. The output dataframe can be interpreted as for any cell, row variable correlation with the column variable is the value of the cell. As mentioned earlier, that the correlation of a variable with itself is 1. For that reason, all the diagonal values are 1.00.

一張含有 文字, 計分板 的圖片

自動產生的描述

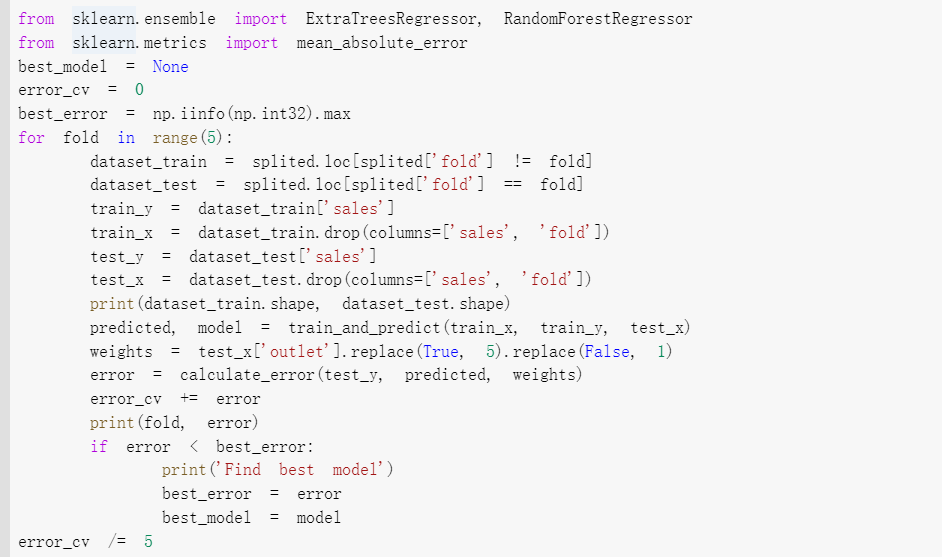
According to the heatmap graph above, heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The more extreme the color (close to all white, all black) grid in the heat map represents the stronger the correlation such as (department\_identifier and sell\_price), (sell\_price and sales) . There is a diagonal all-white line in the heat map such as (sales and sales), (sell\_price and sell\_price). These phenomena should be ignored, because it represents that the relationship between the representation and itself is naturally 1.

* **sklearn.model\_selection.KFold**

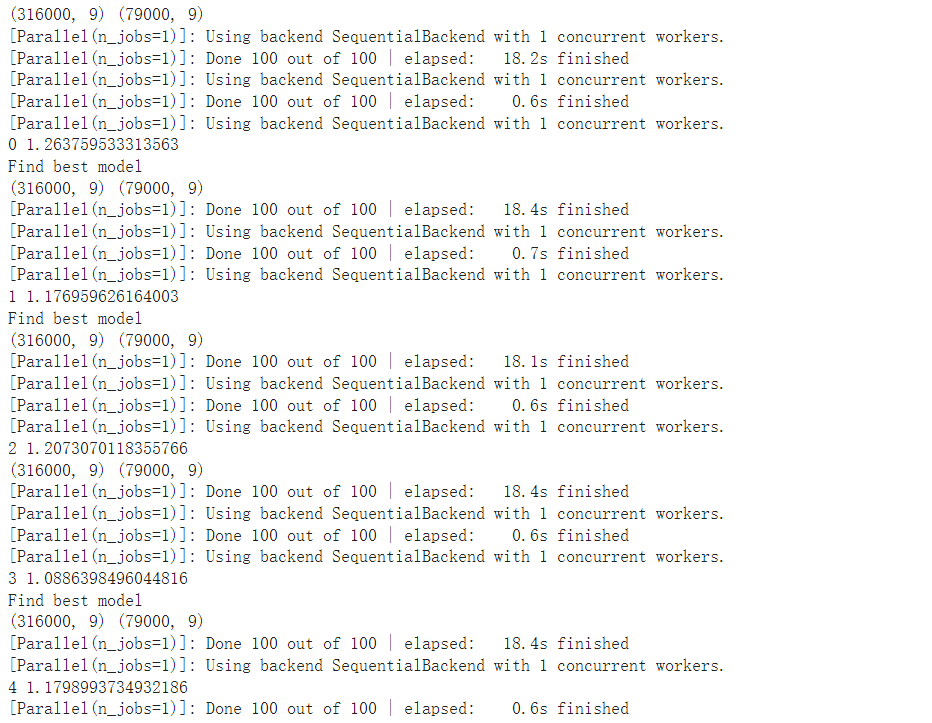
一張含有 文字 的圖片

自動產生的描述

Thirdly, we understood the K-Fold Cross Validation is used to validate your model through generating different combinations of the data you already have. For example, if you have 100 samples, you can train your model on the first 90, and test on the last 10. Then you could train on samples 1-80 & 90-100, and test on samples 80-90. Then repeat. This way, you get different combinations of train/test data, essentially giving you ‘more’ data for validation from your original data. The advantage of K-Fold Cross Validation is intended to avoid the possible bias introduced by relying on any one division into test and train components, is to partition the original set in several different ways and to compute an average score over the different partitions. In our case, we split the data set into 5 parts and try to minimize the deviation of training and test data.



Then, the prediction system would train the system. Each group of fold data would be a train data, and other group of fold would be a test data. The error quantity of each training would be counted. After 5 times training, the minimum error model would be a best model. The best model would be used to predict the sales over a defined period of time.

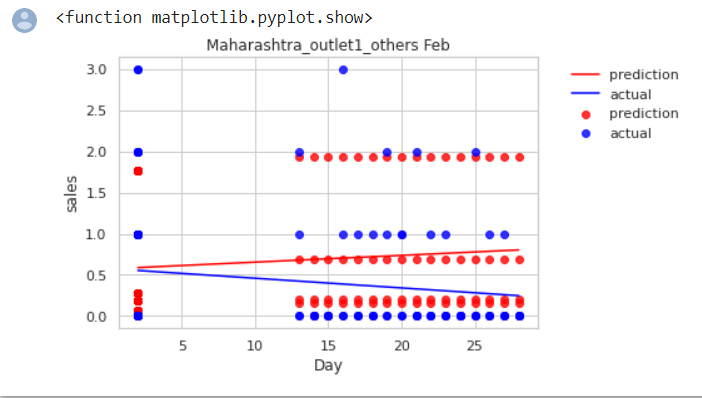


In this case, 1.0886398496044816 is the minimum number, therefore the model 3 would be a best model.

* **sklearn.linear\_model.LinearRegression**

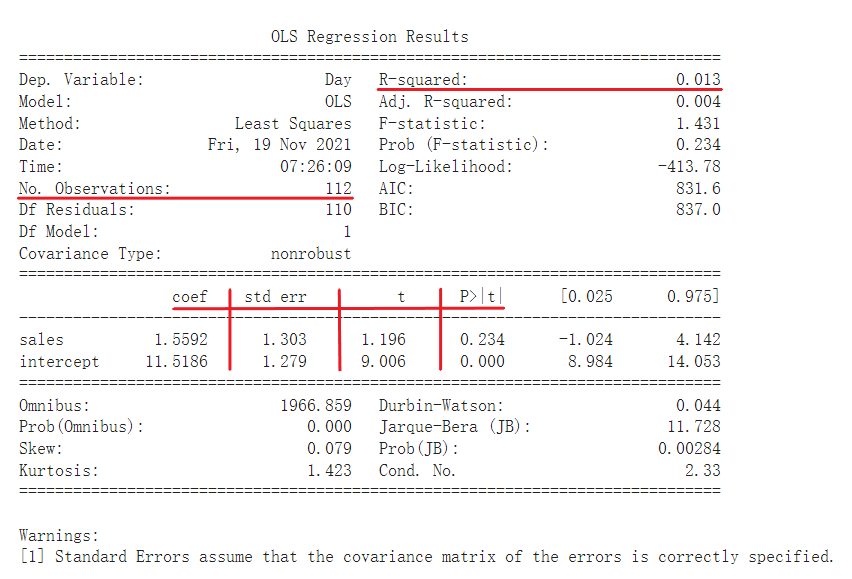
一張含有 文字 的圖片

自動產生的描述



Furthermore, we understood what linear regression is. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). In our example, we wish to estimate the sales of all stores in the city of Maharashtra in February, Therefore, we imported the test\_data\_ANS excel file and defined the variable “x” to get the data location of “Day” to be the x-axis and defined the variable “y” to get the location of “sales” to be y-axis. Two linear regression lines can be seen on the graph: one for the actual line and one for the anticipated line. Every day in February, the genuine linear regression line seemed to be heading downward. On the other hand, the anticipated linear regression line seemed to be increasing.

* **statsmodels.regression.linear\_model.OLS**



According to the graph above, we generated the Ordinary Least-Squares (OLS) Regression which is probably the most used technique in Statistical Learning. We examine the number of 122 data, it can see that the R – squared from the predicting linear regression line equals 0.013, that is indicating a tiny minority data point be close to the linear regression line, therefore it can be noticed that the accuracy of data prediction is not excellent.

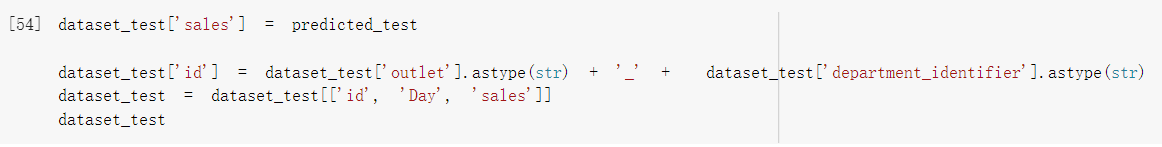
The first column is coefficients that minimize the error function. The coefficient of determination R2 is defined as a ratio of "explained" variance to the "total" variance of the dependent variable y, in the cases where the regression sum of squares equals the sum of squares of residuals.

The second column is standard error. The standard error of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the statistic is the sample mean, it is called the standard error of the mean.

The third column is t-statistic. n statistics, the t-statistic is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error. It is used in hypothesis testing via Student's t-test. The t-statistic is used in a t-test to determine whether to support or reject the null hypothesis. It is very similar to the Z-score but with the difference that t-statistic is used when the sample size is small or the population standard deviation is unknown. For example, the t-statistic is used in estimating the population mean from a sampling distribution of sample means if the population standard deviation is unknown. It is also used along with p-value when running hypothesis tests where the p-value tells us what the odds are of the results to have happened.

In fourth column, the p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

* **The final output**



一張含有 桌 的圖片

自動產生的描述 一張含有 桌 的圖片

自動產生的描述

一張含有 桌 的圖片

自動產生的描述

In the end we got a final output, first of all we need to know what p value is. The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. According to the graph 3 above, we can see that if the p value < 0.001, so it is very strong evidence against the null hypothesis in favor of the alternative. Unfortunately, we can see that overwhelming majority of data are greater than 0.1, only a small minority of data are less than 0.001. It can be seen that most of the data lack evidence against the null hypothesis, so the probability of error is higher.

# Summary

Base on this study case, we understood using Linear regression on inventory demand forecasting would be the most suitable. Though the result is not perfectly predicted, it could satisfy our expectation due to the minor and acceptable std error. Output result is separated in time, which x-axis is within a month, and in product type or id. From these results, we can assume that this demand forecasting system can help us to handle the rough demand and manage the product effectively.

For our study in machine learning, we understood more about sklearn models, and how it can be implemented in our life. We found a variety of uses in dataframe and the visualizing figure is good for us to develop the progress.

Last, we would like to enhance our understanding and discover more on this topic. This research brings our interest on machine learning and we would like to take a follow up experiment on the similar topic.

***(2432 words)***

# References

1. 8 Best Inventory Demand Forecasting Techniques - EazyStock  
   <https://www.eazystock.com/blog/8-best-inventory-demand-forecasting-techniques/>
2. What is Inventory Forecasting? | Definition, Methods & Formula  
   <https://www.zoho.com/inventory/guides/inventory-forecasting.html>
3. Predicting the sales of products of a retail chain - Kaggle  
   <https://www.kaggle.com/mragpavank/predicting-the-sales-of-products-of-a-retail-chain>
4. Why demand forecasting is importance

<https://aimpointdigital.com/importance-of-demand-forecasting/>

1. Demand forecasting problems

<https://www.veryableops.com/blog/demand-forecasting-problems>

1. Ordinary Least Squares in Python

<https://www.datarobot.com/blog/ordinary-least-squares-in-python/>

1. Understanding the P-Value in Regression.

<https://medium.com/analytics-vidhya/understanding-the-p-value-in-regression-1fc2cd2568af>

1. The p-value explained

<https://www.scribbr.com/statistics/p-value/>

1. Python | Pandas dataframe.corr()

<https://www.geeksforgeeks.org/python-pandas-dataframe-corr/>

1. [Day29] Machine Learning: Cross-Validation!

<https://ithelp.ithome.com.tw/articles/10197461>

1. A Beginner’s Guide to Linear Regression in Python with Scikit-Learn

<https://www.kdnuggets.com/2019/03/beginners-guide-linear-regression-python-scikit-learn.html>

…

**ITP4514 Group Assignment Submission Checklist**

**Before the submission of my work, I (the signed party) admit that:**

*(Please tick the corresponding box.)*

Yes No

1. I do not ***DIRECTLY COPY*** from any internet resources. c c
2. I do not copy from **other students’** work. c c
3. I have **summarized/paraphrased** my work from different c c

kinds of materials.

1. I have fulfilled the word limitations (i.e. exceeded the c c

***minimum required words***).

1. My work contains ***more than one paragraphs***. c c
2. With all my works that I have referred from different c c

sources, I have performed ***sufficient referencing.***

1. **If any violations on the above rules are found,** c c

**I acknowledge that my marks might / would be**

**deducted in certain proportion.**

**\_\_\_2021/11/19\_\_**

**(Date)**

|  |
| --- |
| Lo Ching Wa |
| Li Chi Kit |
| Luk Chin Pang |
| Ho Chun Man |

**(Signature)**