PREDICT PRICE SALE FORECASTING

Contents

[INTRODUCTION 1](#_Toc87838132)

[Problem statement 2](#_Toc87838133)

[DATA ANALYSIS 2](#_Toc87838134)

[Exploration 2](#_Toc87838135)

[Data Visualization 4](#_Toc87838136)

[Date wrangling 5](#_Toc87838137)

[IMPLEMENTATION 7](#_Toc87838138)

[Random Forest 7](#_Toc87838139)

[RESULTS 7](#_Toc87838140)

[Accuracy 7](#_Toc87838141)

[Means squared error and root means squared error 8](#_Toc87838142)

[REFLECTION 8](#_Toc87838143)

[REFERENCES 9](#_Toc87838144)

## 

## INTRODUCTION

Sales forecasting is a process where we use previous sales records to predict short term or long term price sales for any business, It has been defined as a pillar in many business as they avoid unforeseen circumstances that may affect business price sales in the future .Furthermore it helps the business to maximize profit by exploiting possible niche that may arise in the future such as Which product will be on demand, Which location the business have vast clients etc. [1]

Sales forecasting is done by implementing a supervised machine learning model where we used historical data and train a model to perform predictions.

We shall use a supervised machine learning model and the algorithm is Random forest classifier

### Problem statement

The goal is to create a machine learning model which predicts the sales over a period of time using historical data. The task involved include the following ‘

* Download a sample data from kaggle site
* Train a ,model test data
* Design a classifier models and use to train data

## DATA ANALYSIS

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. Indeed, researchers generally analyze for patterns in observations through the entire data collection phase (Savenye, Robinson, 2004). Analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

### Exploration

**Train data**



Fig 1 Loading train data into kernel

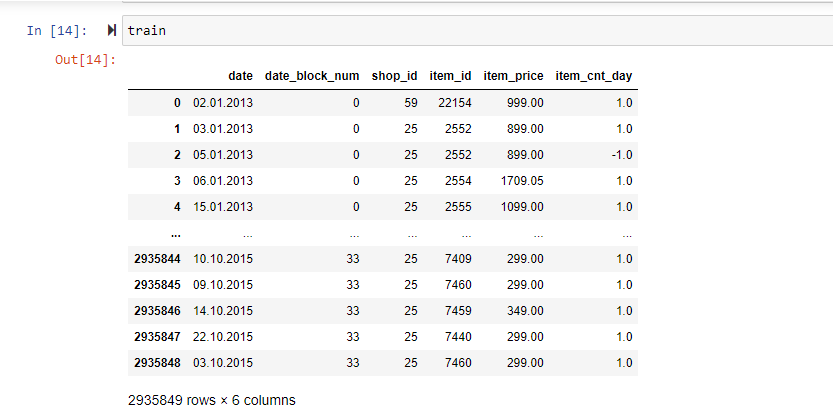


Fig 2: General view of the train data

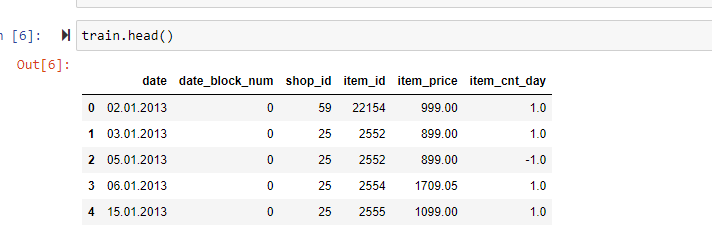


Fig 3: viewing the top head column of the data.

From the figure we can deduce the attributes of the data like

Date->the data purchase are made

Block number

Shop\_id->describing the shop number in data

Item\_id->unique key of the item purchases

Item\_price-> price of the item

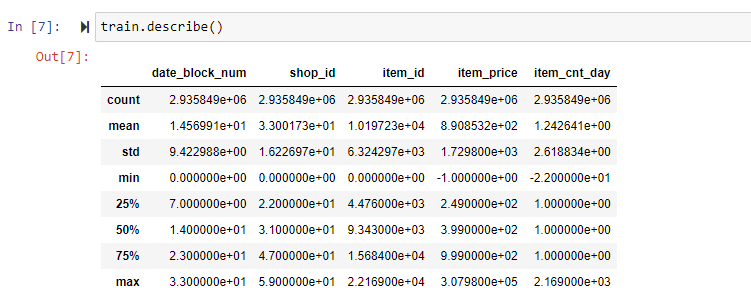


Fig 4 Showing describing the total details of the train data

### Data Visualization

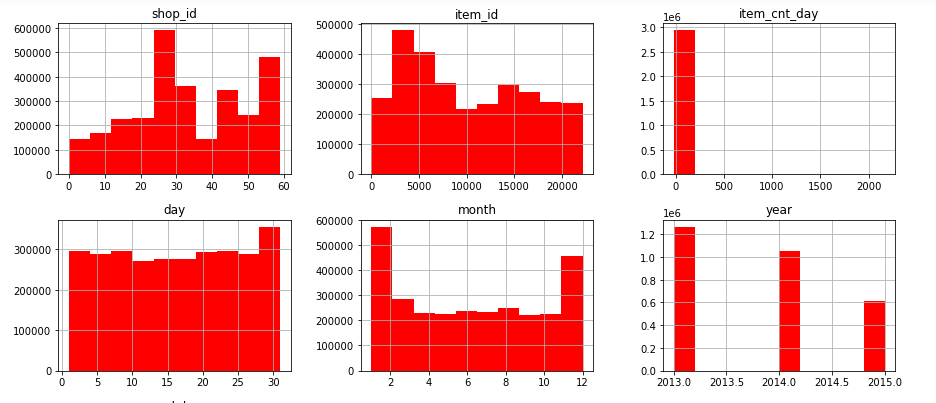


Fig 5: Displaying the total of sales on every column in the train data. The Y axis represent the amount of sale done.

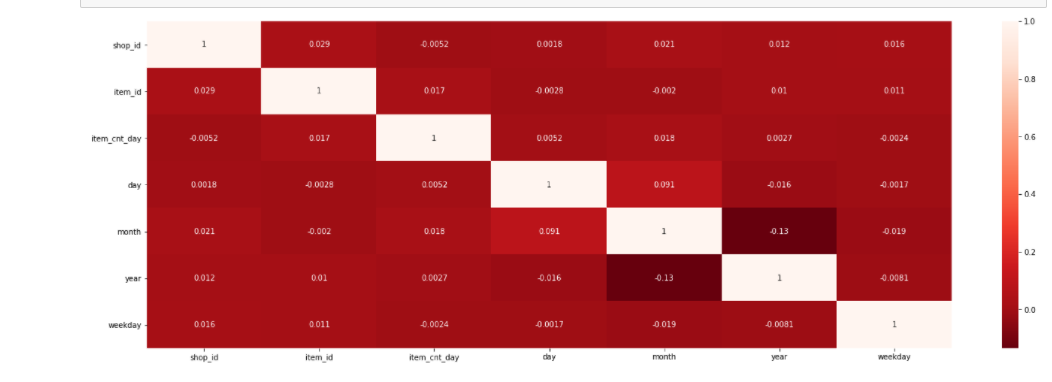


Fig 6: Heat map display, from here we can determine the weight of each attribute so that we can prune the data and use attributes which directly the sales per data.

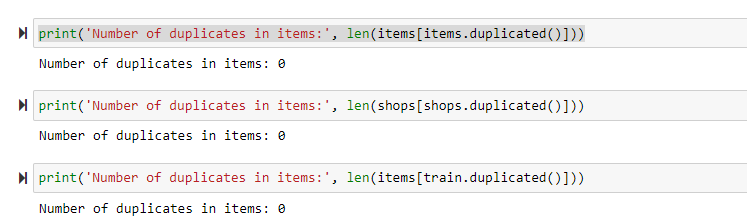
### Date wrangling

* Check for NA and missing values

|  |  |
| --- | --- |
|  |  |
|  |  |

Fig 7: we can deduce that all the data types are on integers which is essential for machine learning model. Furthermore there are no missing values hence we can proceed to create Machine learning model for predicting sales.

* Check for duplicates



The next step is to deal with the date and split into single entities such as day, month and year. This will give a good predictability on a specified time.

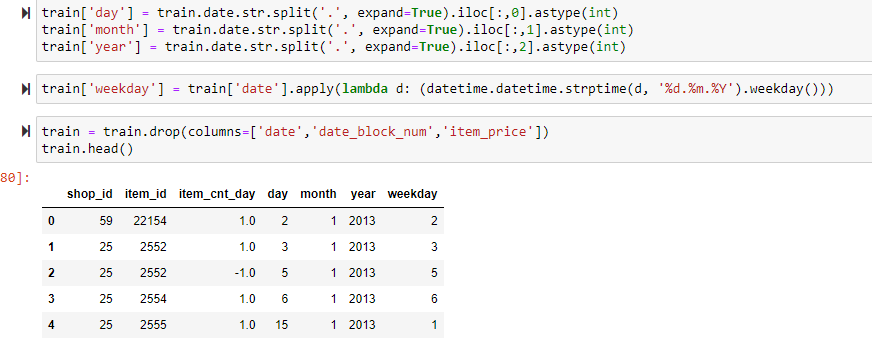


Fig 8:

From the above display we can see we have managed to split the date column into individual entities

## IMPLEMENTATION

The prediction of the price was predicted using supervised learning

The algorithm used is Random Forest with 50 trees

### Random Forest

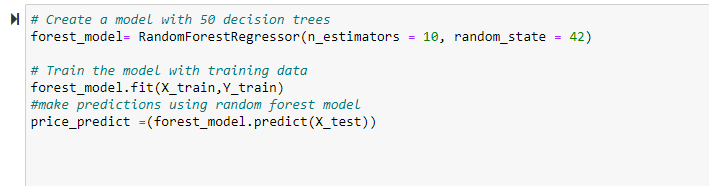


Fig 9

The scikit learn module has inbuilt function for random forest repressor,,we fit the splited module and make prediction using the X\_test.The expected output is the date at which items are predicted to have high sales.

**Advantages of random forest**

1. Reduces over fitting in decision trees improving accuracy
2. Flexible to handle any regression or classification problem
3. Uses rule based approach hence normalization of data is not necessary

**Disadvantages of random forest**

1. Requires more computational power
2. Take time to train the classifier

## RESULTS

### Accuracy

The accuracy of the model refers to the total instances that were correctly predicted. This being a regression model the accuracy is calculated using MSE and RMSE

### Means squared error and root means squared error

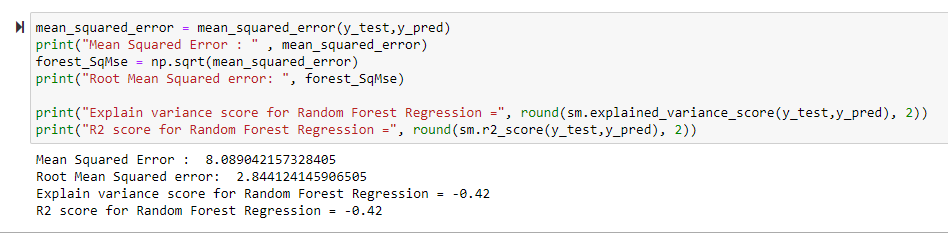


Fig 10

MSE is used to calculate how close the estimates (y\_pred) are close to the actual values. The MSE is 8.089 .The lower MSE means the values are close to the actual

## REFLECTION

The process of design, implementation and result analysis can be summarized as follows

* An initial problem statement was defined and relevant data sources we found at kaggle
* The data was download and segmented to fit the desired learning for the model.
* The model of choice was implemented using Random forest classifier. This gave me chance to explore others algorithm implementation and reason as to why I settled in random forest can be summarized from its advantages

The project was interesting as at the end I was able to implement all theoretical aspect of Ml learned in class. I understood the mathematical behind all ML terms and metric used to measure the score of the model.

I had to familiarize myself jupyter notebook since I was novice user on the same.

The challenge I encountered was that the data I used was a bit large hence the running time for the code was a bit slow.

## REFERENCES

1. https://www.nytimes.com/1981/06/16/business/market-place-the-diverse-art-of-forecasting.html