**INM701**

**Introduction to Artificial Intelligence**

**Coursework Report**

*By*

Vrushang Bawne (Student Number: 210015700)



**MSc Artificial Intelligence 2021-22**

**[Term 1]**

PREDICT PRICE SALE FORECASTING

Contents

[INTRODUCTION 2](#_Toc94651894)

[Problem statement 3](#_Toc94651895)

[DATA ANALYSIS 4](#_Toc94651896)

[Exploration 5](#_Toc94651897)

[Data Visualization 7](#_Toc94651898)

[Date wrangling 7](#_Toc94651899)

[IMPLEMENTATION 9](#_Toc94651900)

[Random Forest 10](#_Toc94651901)

[RESULTS 11](#_Toc94651902)

[Accuracy 11](#_Toc94651903)

[Means squared error and root means squared error 11](#_Toc94651904)

[REFLECTION 11](#_Toc94651905)

[REFERENCES 12](#_Toc94651906)

## 

## 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

A Sales Forecast permits organizations to foresee their feasible deals, center their assets in a more financial way, and plan their future improvement all the more unequivocally. To make the estimate more exact, the organization ought to likewise watch out for monetary patterns that will generally change rapidly, and industry correlations.

There are various techniques for making a Sales Forecast and every last one of them utilizes a particular Sales Forecast recipe. Despite the fact that the recipes utilize various conditions they give comparable outcomes with a slight level of incorrectness. No matter what the sort of deals anticipating demonstrating an organization picks, they will by and large get pretty comparable outcomes, assuming the computations and numbers are exact. All things considered, assuming the organization is battling with picking the best Sales Forecast strategy they ought to talk with a broad pragmatic expert information in this field.

Subsequent to investigating the information, it ought to be organized into a report. To make it as careful, simple to peruse, and useful as conceivable it ought to contain the accompanying data:

**Data about the organization**  :- To express the points of interest of the report, it should begin with a title. It ought to be trailed by the organization's name and area. For few organizations set their logo on the reports to give a more interesting plan.

**Item data** :- Here the report should assign data about the item. It can incorporate its name, depiction, principle attributes, and so forth This piece of the figure ought not be over-burden with data since gauges are typically devoted to a few sorts of items.

**The period of the forecast**  :- By and large, figures are made to foresee deals for the following a year, nonetheless, the time span of the archive can be changed relying upon the requirements of the organization. It ought not be too reached out later on however, since it is difficult to foresee what the monetary inclinations will resemble. It can influence the exactness of the estimate and it won't fill its primary need.

**Past deals information** :- This archive is generally founded on data about past deals. It is up to the organization how to introduce it, in any case, the most well-known method for doing it is in a type of a timetable. Every long stretch of the earlier year should be remembered for it, just as the benefit sum made during this particular time span.

**Connections**:- To exhibit the examination between the time span or between the deals of various items the organization can utilize realistic materials, like outlines, graphs, plans, and so on For them has many advantages for the organization, for example, they help to comprehend the materials better, set the attention on the fundamental parts of the estimate, and add uniqueness to the archive by outwardly communicating the information of the gauge.

### 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

You are provided with daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

File descriptions

sales\_train.csv - the training set. Daily historical data from January 2013 to October 2015.

test.csv - the test set. You need to forecast the sales for these shops and products for November 2015.

sample\_submission.csv - a sample submission file in the correct format.

items.csv - supplemental information about the items/products.

item\_categories.csv - supplemental information about the items categories.

shops.csv- supplemental information about the shops.

## 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.

Data fields

ID - an Id that represents a (Shop, Item) tuple within the test set

shop\_id - unique identifier of a shop

item\_id - unique identifier of a product

item\_category\_id - unique identifier of item category

item\_cnt\_day - number of products sold. You are predicting a monthly amount of this measure

item\_price - current price of an item

date - date in format dd/mm/yyyy

date\_block\_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33

item\_name - name of item

shop\_name - name of shop

item\_category\_name - name of item category

This dataset is permitted to be used for any purpose, including commercial use.

### Exploration

**Train data**



Fig 1 Loading train data into kernel

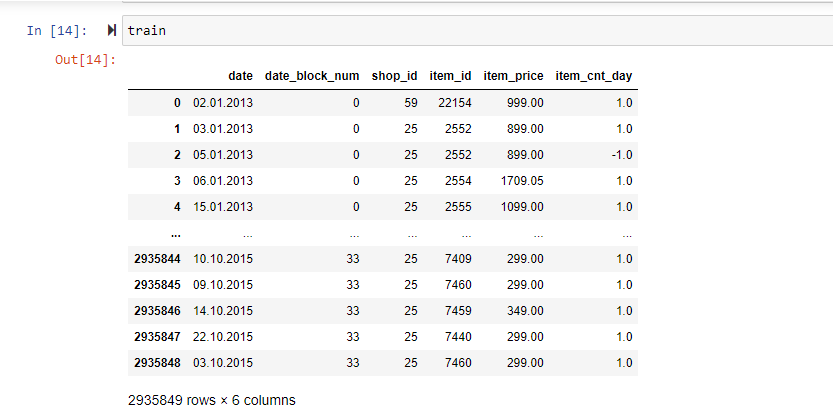


Fig 2: General view of the train data

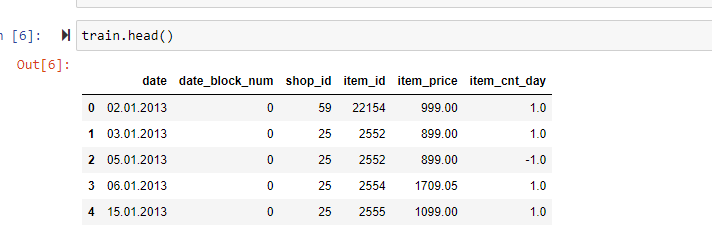


Fig 3: viewing the top head column of the data.

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

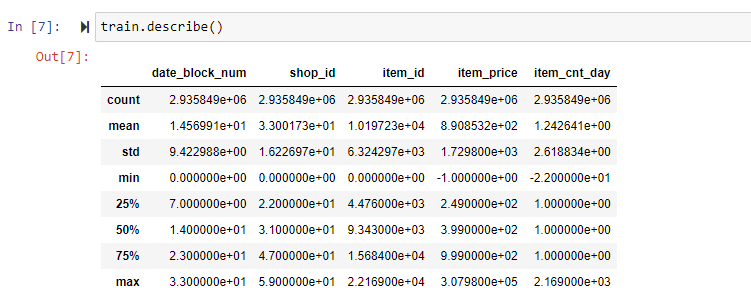


Fig 4 Showing describing the total details of the train data

### Data Visualization

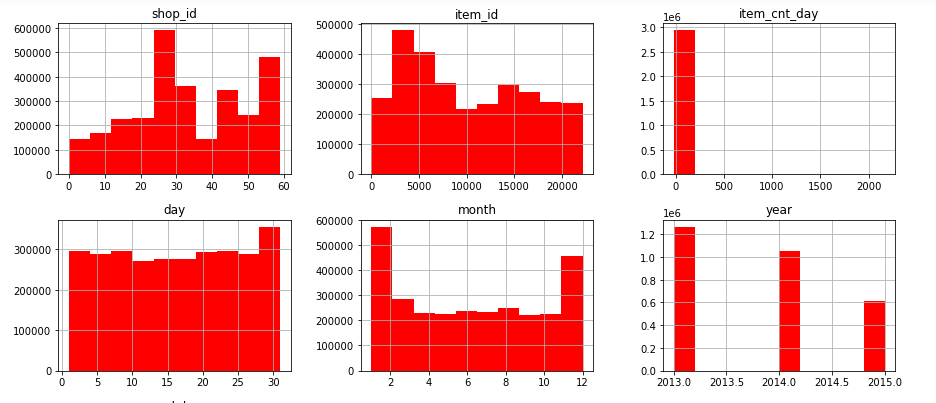


Fig 5: Displaying the total of sales on every column in the train data. The Y axis represent the amount of sale done.

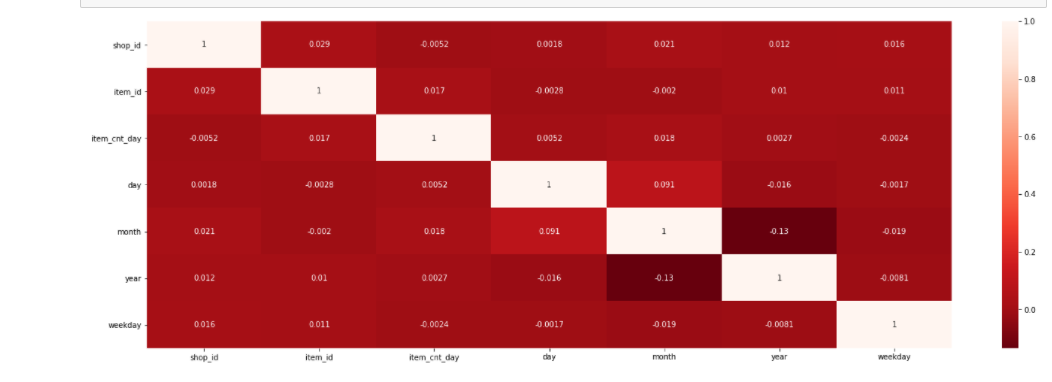


Fig 6: Heat map display, from here we can determine the weight of each attribute so that we can prune the data and use attributes which directly the sales per data.

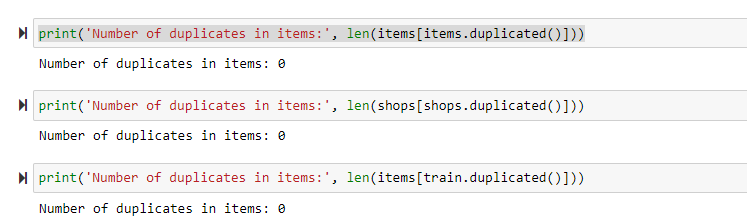
### Date wrangling

* 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.

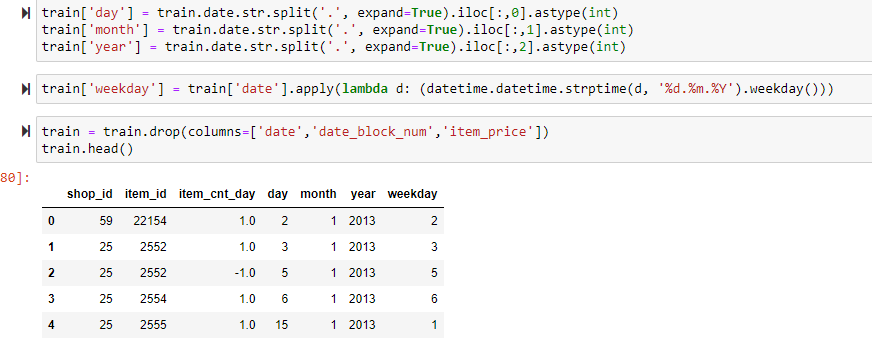


Fig 8:

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

## IMPLEMENTATION

Tree models are known to be high fluctuation, low inclination models. In result, they are inclined to overfit the preparation information. This is appealing in the event that we restate how a tree model treats we don't prune it or present early halting standards like a base number of examples per leaf hub. All things considered, it attempts to divide the information along the elements until the occurrences are unadulterated with respect to the worth of the objective component, there are no information left, or there are no highlights passed on to spit the dataset on. On the off chance that one of the above remains constant, we grow a leaf hub. The outcome is that the tree model is developed to the maximal profundity and therewith attempts to reshape the preparation information as exact as conceivable which can without much of a stretch lead to overfitting. One more downside of old style tree models like the (ID3 or CART) is that they are moderately unsound. This shakiness can prompt what is happening that a little change in the structure of the dataset prompts something else entirely model.Random forest is an ensemble of decision tree algorithms.

The prediction of the price was predicted using supervised learning

The algorithm used is Random Forest with 50 trees

### Random Forest

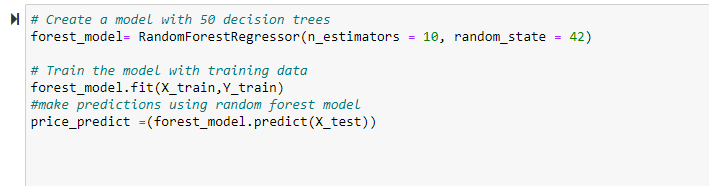


Fig 9

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
4. One of advantages of Random Forest which exists me more than anything else is, the force of handle huge informational collections with higher dimensionality. It can deal with huge number of info factors and character most huge factors so it is considered as one of the dimensionality decrease strategy. Further, the model results significance of variable, which can be an extremely convenient element.
5. Random forest can solve both type of problems that is classification and regression and does a decent estimation at both fronts.

**Disadvantages of random forest**

1. Requires more computational power
2. Take time to train the classifier
3. It definitely works really hard at order however not with respect to relapse issue as it doesn't gives exact consistent nature forecast. If there should arise an occurrence of relapse, it doesn't foresee past the reach in the preparation information, and that they may over fit informational collections that are especially boisterous.
4. Random forest can feel like a black box approach for a statistical modelers we have very little control on what the model does. You can at best try different parameters and random seeds.

## 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

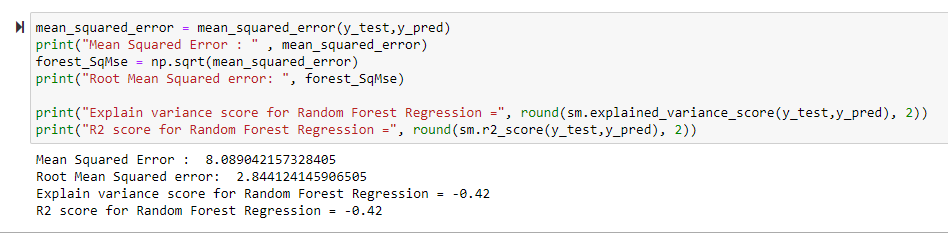
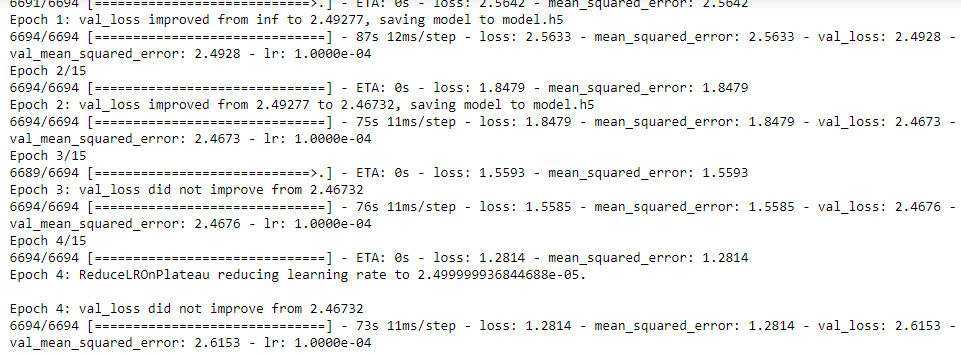


Fig 10

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

# LSTM

As an alternative model, I tried to implement a LSTM network and trained it on our dataset. We have created LSTM layers. The network takes the input as the number of features on which the model is trained. For the output, we have a softmax layer which classifies the data as one of the two possible classes We run the LSTM for 15 epochs. and set the verbose = 2 to see the detailed performance of training. We can see that the loss function is high at the start and keeps minimizing with the epochs.



## REFLECTION

First I downloaded a sample data form kaggle site then trained the model after that Design a classifier model and used to train model data. Then I loaded the data into kernel the output was general view of the data. Used Data visualization to display the total of the sales on every column in the data. Y axis represent amount of sales done and also the Heat map Display from that we can determine the weight of each attribute so that we can prune the data and use attributes which directly the sales per data .then we used Data wrangling to check the Na and missing values. Used random forest with 50 decision trees. Accuracy of the model refers to the total instances that were correctly predicted.

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.

My contribution helped make the project a success, by encouraging to brainstorm, research, and provide inputs and ideas as well. I spent a considerable amount of time, But eventually, I came up with a logical explanation and considered all opposing views. I went out of my scope of work to try and get a better glimpse of how the task can handle efficiently to sure the highest score possible.

I had to familiarize myself jupyter notebook since I was novice user on the same. by laying a plan which will ensure every participant is involved and the time factor is kept at a glance. I did manage to complete the assignment in the specified time and everyone did their task and research as expected. Even when there were opposing views, I would always harmoniously settle the view. Furthermore, I have provided motivation and enthusiasm, which has been essential to the success of the project

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.

# Conclusion

According to the aim, we wanted to classify the insurance claims made as fraudulent or sincere.

Initially, our model had accuracy of . The MSE is 8.089%. This was however not very effective

as the dataset was skewed and the model couldn’t learn about tuples belonging to class 1 but still

managed to get a high accuracy. So, for this purpose we reduced the dataset and then tried again.

Here we saw a drop in the accuracy, but the model was able to learn class 1 tuples.

LSTM was implemented as the alternative model. We tried hyperparameter tuning and saw the

performances of the different models.

## REFERENCES

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

GITHUB LINK:-

https://github.com/vrushangb/INM701-Vrushang