Table of Contents

[1. Introduction 2](#_Toc78137222)

[2. Business Understanding: Defining the Problem 3](#_Toc78137223)

[Defining the Target Variable 3](#_Toc78137224)

[3. Data Understanding 3](#_Toc78137225)

[4. Data Preparation 5](#_Toc78137226)

[Exploratory Data Analysis (EDA) 6](#_Toc78137227)

[Text preprocessing 7](#_Toc78137228)

[Visualization 8](#_Toc78137229)

[5. Modeling 12](#_Toc78137230)

[Model Evaluation 14](#_Toc78137231)

[6. Deployment 15](#_Toc78137232)

[7. Summary and Conclusion 15](#_Toc78137233)

[**Next Steps** 16](#_Toc78137234)

[8. References 17](#_Toc78137235)

## Introduction

Retail Data Analytics is important process of analyzing data to make important decisions which help to improve the sales, customer satisfaction, quality.

Retail Data Analytics helps to supply chain and inventory management such as how many products left, stock details etc.

Benefits of Retail data analytics [ Source – Ref 3]

* Customer satisfaction
* Customer retention
* Effective Operative procedures for inventory and labor [ Source - Ref 4]

In the recent competition, The Owner of Shop should study the behavior of customers buying pattern. In that way we can advertise and improve the needs for better sales and customer service.

In this project, I am focused on predicting the sales of shop and departments in future based on the forecasting and performance of sales. Also, especially important correlation between sales and Department, temperature, Holidays, and other parameters.

Without predictive model, the only way to be sure to reach all of them is to reach every single customer. Instead of that, we can build accurate predictive model and we can target much more effective way.

The below are the reasons I have selected Python Language to perform this project.

* Python as a programming language has numerous uses such as Machine Learning, AI, mobile applications, statistical , etc. As we know, AI, machine learning, and data analysis/analytics is where it has amassed most of its popularity.
* Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc. Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc.

## Business Understanding: Defining the Problem

The business objective remained the same: The main goal of the project is to build a predictive model to predict or determine the sales of the shops and Department and then share the recommendation to improve Sales. Also, to identify the correlation between sales and Department, temperature, Holidays, and other parameters. I am planning to use different classification machine learning other classification models such as SVM, Logistic Regression and Time Series Forecasting's identify the Ensemble Classifier had the best accuracy.

## Data Understanding

The format of the dataset for this case study is CSV. The data contains below variables.

Data used in this project are online product reviews collected from Kaggle.com ( Women’s Clothing E-Commerce dataset )

https://www.kaggle.com/manjeetsingh/retaildataset

**Tools and Library to be used in this Project**

* Pandas library used to load the data from csv to Dataframe.
* Seaborn and matplotlib are used for data visualization.
* Sklearn library used for model and data preparation.



The format of the dataset for this Project is CSV. The data contains below variables.

Pandas library used to load the data from csv to Dataframe.

**Features Dataset**

Contains additional data related to the store, department, and regional activity for the given dates.

1. Store id

2. Date

3. Temperature

4. Fuel\_Price

5. Markdowns

**Store Dataset**

1. Date - Sales Date
2. Temperature - Region's Temperature
3. Fuel\_Price - Region's Fuel cost
4. MarkDown1-5 - markdowns
5. CPI - the consumer price index
6. Unemployment - the unemployment rate
7. IsHoliday - whether the week is a special holiday week

**Sales Dataset**

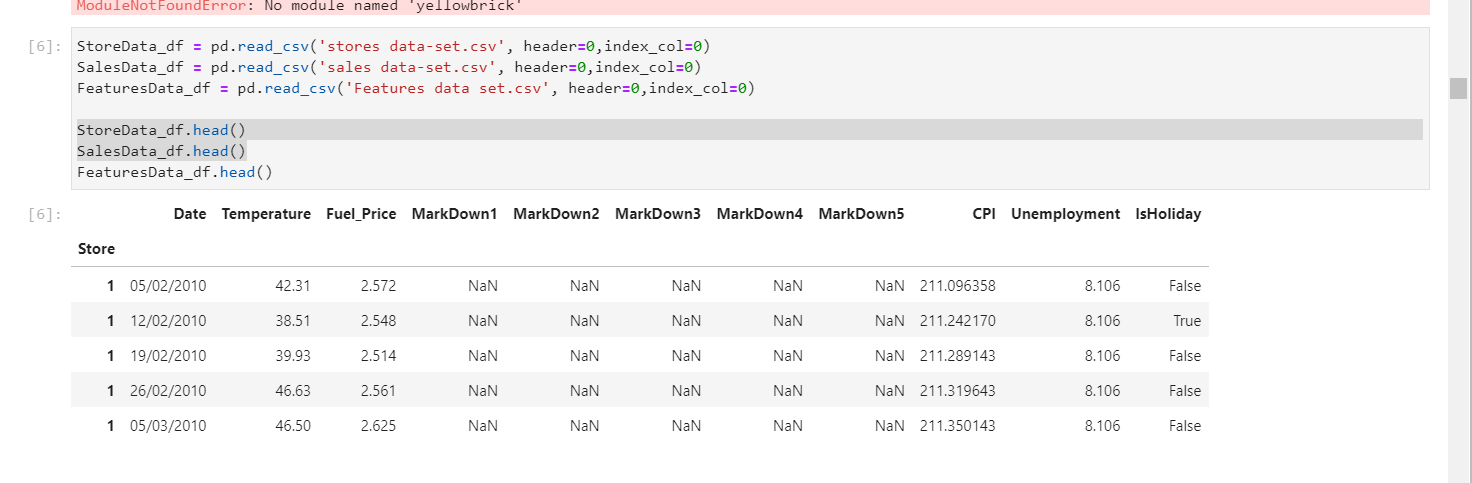
Historical sales data - Feb 2010 to May 2011

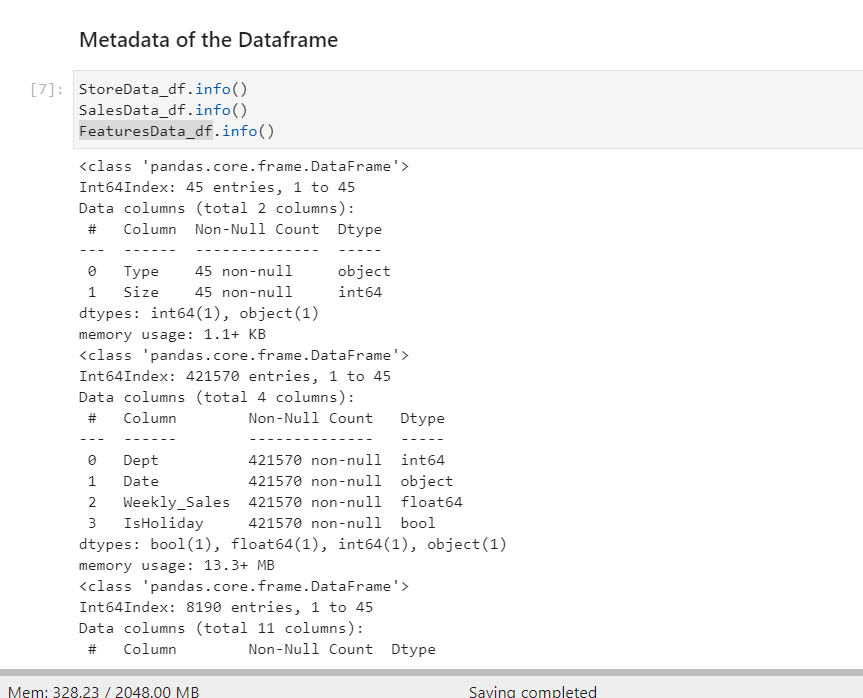
1. Store - the store number
2. Dept - the department number
3. Date - the week
4. Weekly\_Sales - sales for the given department in the given store
5. IsHoliday - whether the week is a special holiday week

## Data Preparation

Based on the reading, “data scientists usually spend 70% of the project time here, preprocessing and exploring the data”. So, the Data Preparation is important for predictive analytics project. We need to make sure the data is suitable and well prepared for model. While preparing the Data, we need to verify the metadata of the given dataset and need to understand very clearly about data patterns, relationships etc. [ Source – Ref 5]

* Need to describe the dataset describe()
* Need to understand the number of columns and rows shape.
* Need to understand the summary of the data ( info )



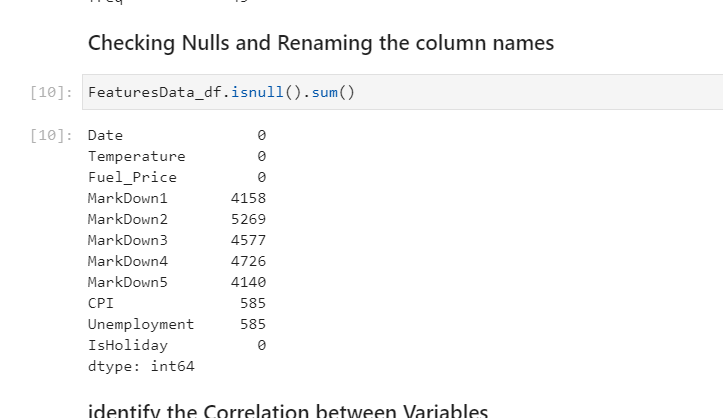


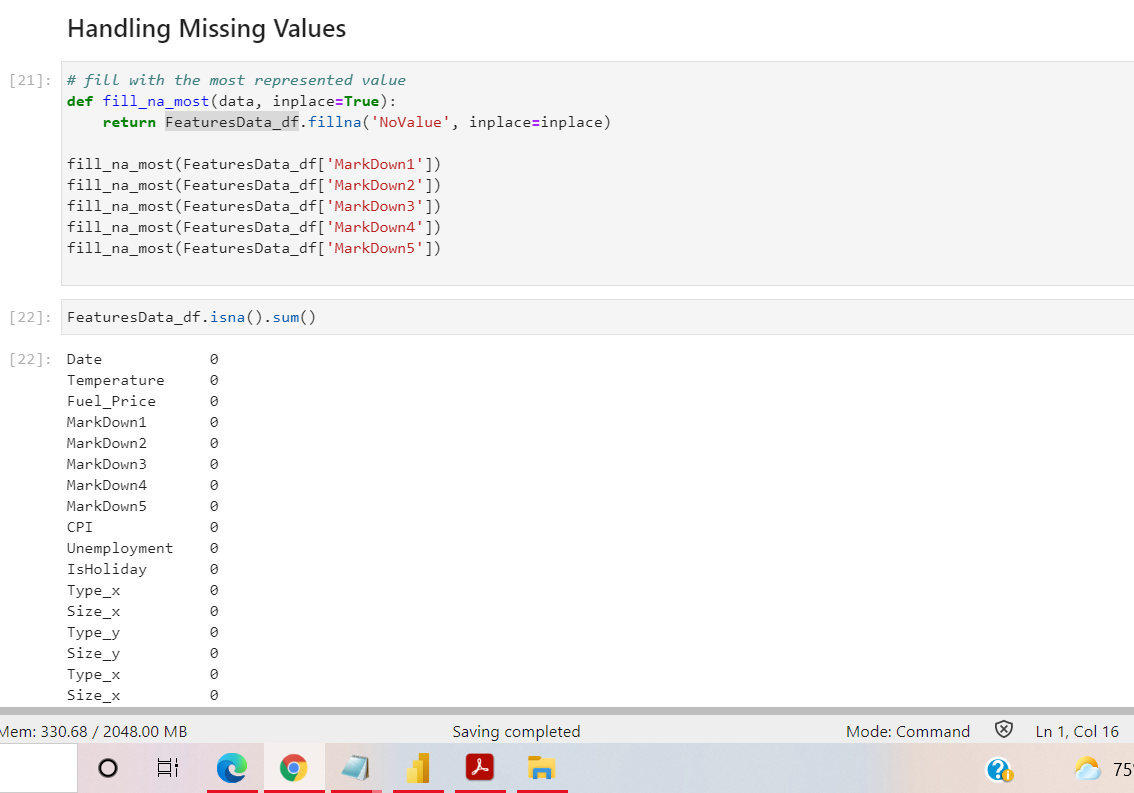
### Exploratory Data Analysis (EDA)

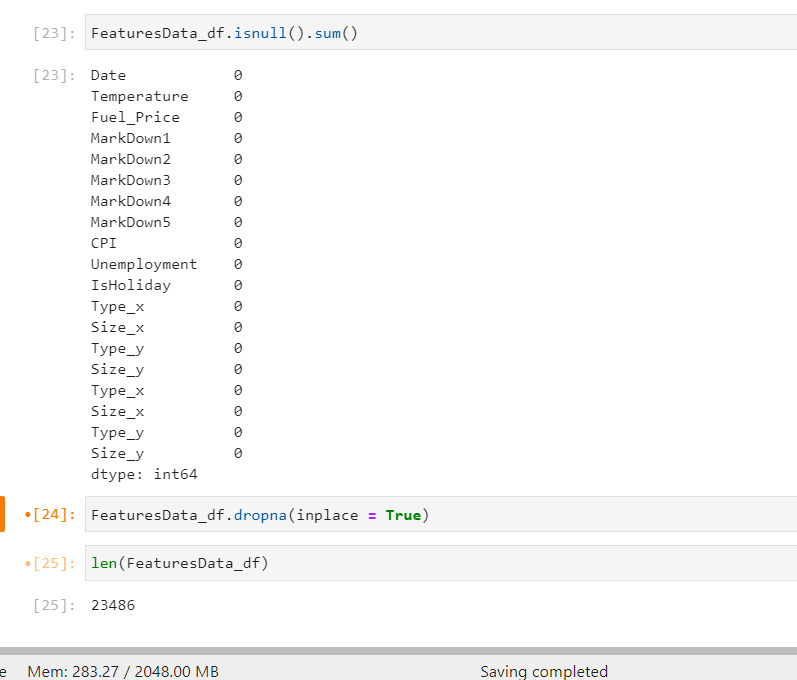
We need to perform EDA and that data needs to be cleaned to gain insights that may be useful in improving the performance of our model. Few of the EDA as follows

* Filling missing information (Null / NA values)
* Encoding of categorical variables
* Dropping of features (or columns)

The above Missing information handled.

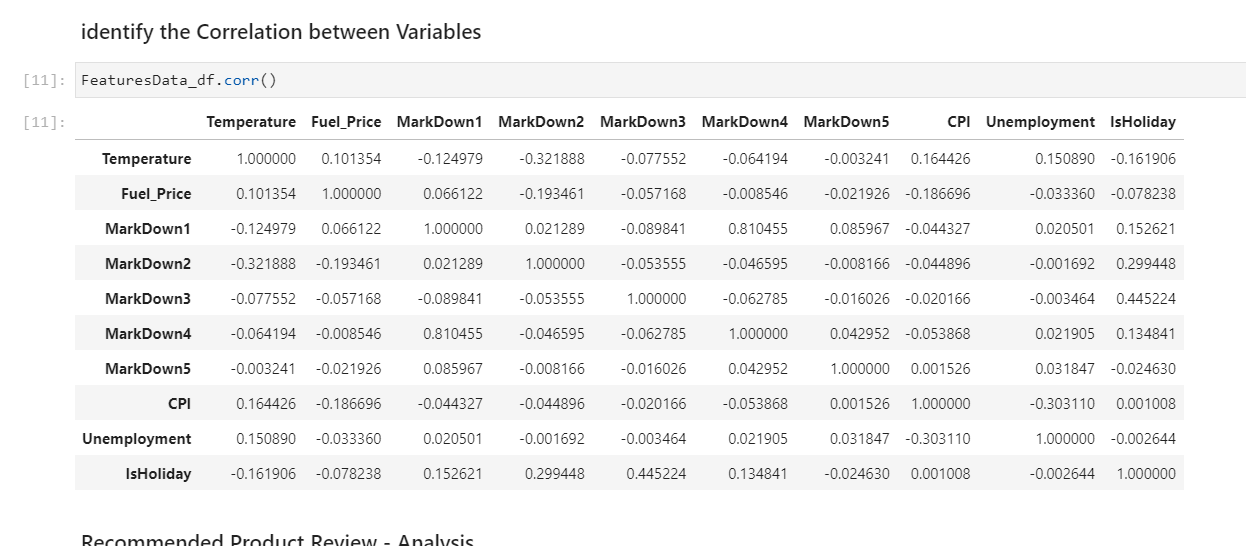


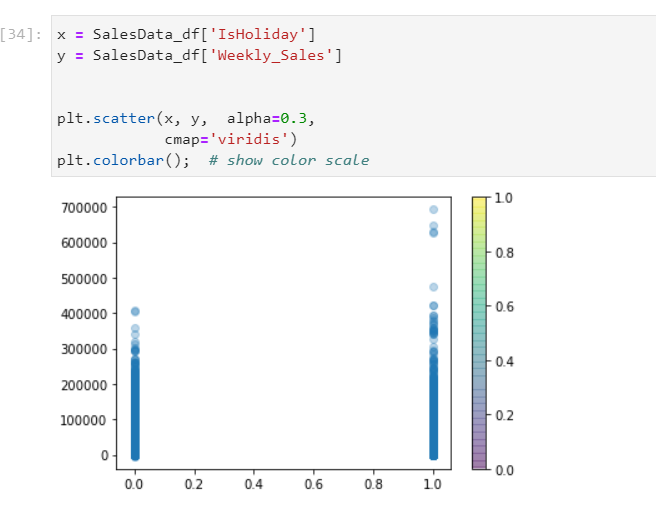




### Visualization

Visualization used to show behavior of data and explains insight of data.

1. Identified the Correlation between variables in the dataset
2. 
3. Analyzed the Holidays vs Non-Holidays Sales



## Modeling

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). Hence in our Case Study we can use Logistic Regression methods to predict the outcome claim as 1 or 0 .

I have used 70%-30% split between training and Testing data.

The below are models verified by training the data.

* Logistic Regression
* SVM
* MultinomialNB
* BernoulliNB
* KNN
* RandomForest
* DecisionTree

## Summary and Conclusion - Milestone 1

* Loaded the data from .csv files and performed the cleaning and verification.
* Performed Exploratory Data Analysis on dataset to produce insights such as trends over the target variables, distribution of counts, Histogram comparison. Removed the missing values, removed the unwanted columns.
  + Handled the NA values and NULL Values.
* Preprocessed the Text data and identified the Positive and Negative comments
* Lowercase
* Removing Punctuation
* Tokenization
* Stopword Filtering
* Stemming
* Split the data set into Training – Testing in the ration of 70% : 30%

### **Next Steps**

This predictive system is built for product review, and we can expand this approach to deal with Product like Amazon, Flipchart by using Product Reviews API data or many social media sites application programming interfaces (API) data ( example twitter / Facebook )

## References

1. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by Eric Siegel

Page number – 166 and 167

1. Applied Predictive Analytics by Dean Abbott ( Principles and Techniques for the Professional Data Analyst )
2. <https://www.stitchdata.com/resources/retail-data-analytics/#:~:text=Retail%20data%20analytics%20is%20the,are%20targets%20for%20data%20analytics>.
3. <https://www.kaggle.com/manjeetsingh/retaildataset>
4. <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=1ffbe4896f63>