Business forecasting using machine learning

MID SEMESTER Report
Submitted in partial fulfillment of the requirements of
CS F376 Design Project
By

Siddharth Choudhury
IDNO: 2020A7PS0028U

Under the supervision of
Dr. Siddhaling Urolagin
Professor

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

DUBAI CAMPUS, DUBAI UAE

NOVEMBER - 2022

ACKNOWLEDGEMENTS

I would like to express my deepest sense of gratitude, first and foremost, to my Supervisor **Dr. Siddhaling Urolagin**, Professor, Computer Science Department, BITS Pilani, Dubai campus, United Arab Emirates, for her valuable guidance and encouragement during the course of this Project. I am extremely grateful to him for his able guidance, valuable technical inputs and useful suggestions.

I express my sincere thanks and gratitude to our Director, BITS Pilani, Dubai Campus, **Prof. Dr. Srinivasan Madapusi,** for their motivation, encouragement and support to pursue my Project.

I am grateful to examiners **Dr. Siddhaling Urolagin** for their valuable suggestions.

Above all, I thank the Lord for giving me the strength to carry out this work to the best of my abilities.

Name : Siddharth Choudhury
ID No. : 2020A7PS0028U

CERTIFICATE

This is to certify that the Mid Semester Project Report entitled, in partial fulfillment **Business forecasting using machine learning** of the requirement of CS F376 Design Project embodies the work done by him under my supervision.

Date:

Signature of the Supervisor

Name: Dr Siddhaling Urolagin
Designation: Professor (CS)

BITS Pilani, Dubai Campus

FIRST Semester 2015-2016

Project Course Code and Course Name:
Semester: First Semester 2022-2023
Duration: 10.09.2022-5.01.2023
Date of Start: 10.09.2022
Date of Submission: 20.11.2022
Title of the Report: Business Forecasting Using machine learning

ID No. / Name of the student: **2020A7PS0028U / SIDDHARTH CHOUDHURY**

Discipline of Student: B.E (Hons.) Computer Science

Name of the Project Supervisor: Dr SIDDHALING UROLAGIN

Key Words: **Machine Learning, Heat maps, Categorial data plots, Correlation plots, Data preprocessing**

Project Area: **Machine Learning**

Abstract: **Business forecasting is critical for retailers since it is necessary for a variety of operational choices. Forecasting demand on special days, when demand patterns are very different from those on typical days, is one of the biggest challenges. We discuss the issue of predicting the daily demand for various product categories at the shop level using the example of a supermarket chain. These projections serve as a guide for purchasing and manufacturing decisions. We address the forecasting issue using machine learning. We describe and talk about the potential of creating a classification problem rather than a regression problem in specific. Machine learning techniques outperform traditional methods empirically, whereas classification-based approaches outperform regression-based approaches. We also discovered that machine learning techniques are better suited for use in a sizeable demand forecasting scenario that frequently happens in the retail sector, in addition to offering more accurate forecasts.**

Signature of Student **Signature of Supervisor**

Date: 31/10/2015 **Date: 31/10/2015**

TABLE OF CONTENTS

BUSINESS FORECASTING USING MACHINE LEARNING

INTRODUCTION

Sales forecasting has always been a very important topic to focus on. All vendors must now anticipate well and optimally in order to maintain the effectiveness of marketing groups.
Manually performing this work would be time-consuming, which is undesirable in today's fast-paced environment and could result in grave mistakes that would result in bad management of the firm. A significant portion of the world economy is dependent on the business sectors, which are literally expected to generate enough goods in the right amounts to satisfy demand.

The primary objective of business sectors is market audience targeting. It is crucial that the business has been successful in achieving this goal by utilizing a forecasting system. Forecasting requires examining data from a variety of sources, including market trends, customer behavior, and other elements. The companies would benefit from this analysis by having better financial resource management. The forecasting method can be used for a variety of things, such as estimating future demand for the product or service or estimating how much of the product will be sold in a specific time frame.

Here, machine learning has a lot of potential for use. In the field of machine learning, computers are able to execute some jobs better than people. They are employed to carry out specific tasks in a methodical manner and produce improved outcomes for the advancement of the modern civilization.
The foundation of machine learning is mathematics, which may be used to design various paradigms that are close to the ideal output. Machine learning has been shown to be beneficial in the instance of sales forecasting. It aids in more precise forecasting of upcoming sales.

In this project report, we suggested machine learning techniques for data gathered from a grocery store's prior sales. Based on a few key characteristics identified from the available raw data, the goal is to predict the sales pattern and the quantities of the products to be sold. To fully understand the data, analysis and study of the acquired data have also been done. At each crucial stage of the marketing strategy, analysis would assist business organizations in arriving at a probable decision.

DATA VISUALIZATION

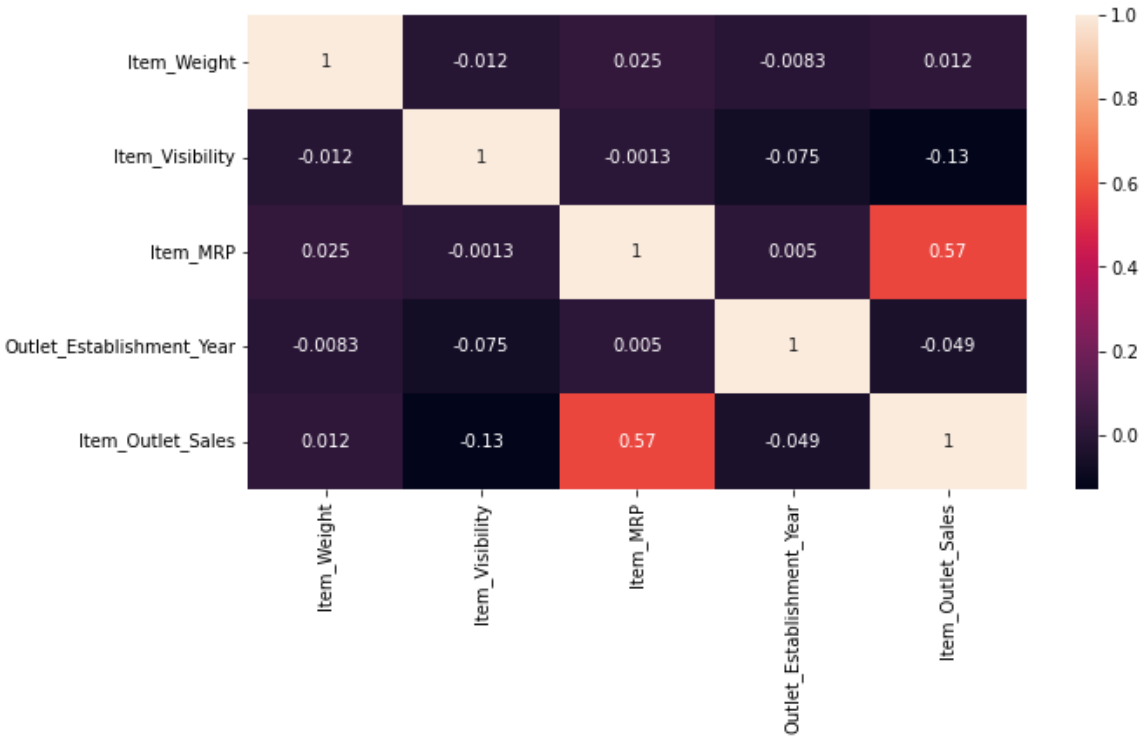
HEAT MAP

Heat map for figuring out the relationships between the dataset's characteristics.

Here, the correlation between the target variable and the other qualities is shown using a heat map, a color-coded matrix from the Seaborn data visualization toolkit.

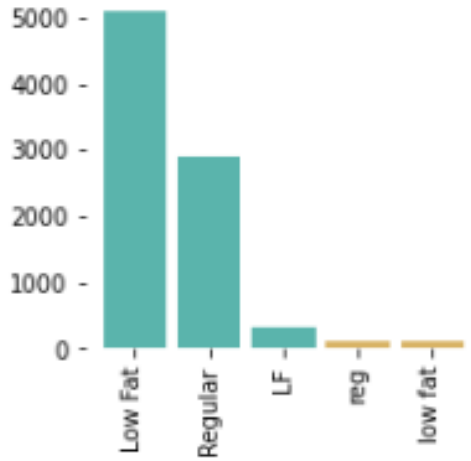
The target variable's dependence on an attribute decreases as the color intensity of the attribute's relative to the target variable increases.

The goal variable, Item_Outlet_Sales, is seen to be most dependent on item MRP and least dependent on Item_Visibility. Therefore, higher the MRP of an item, lower will be the Item_Outlet_Sales.

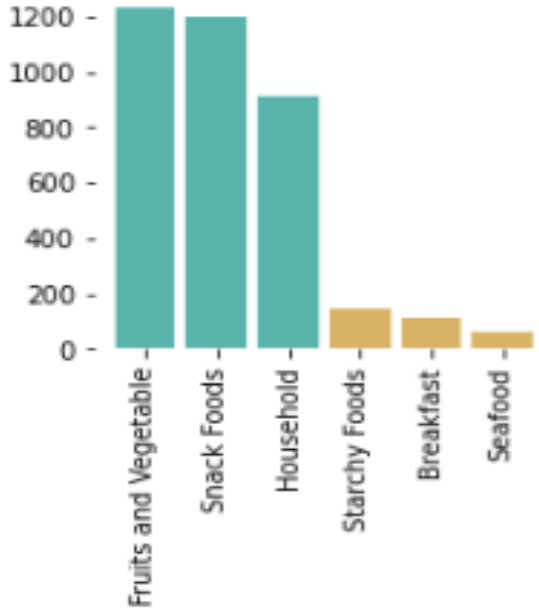


CATEGORIAL DATA PLOTS

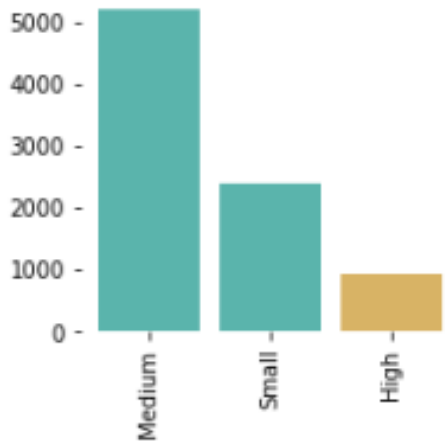
The distribution of various Item fat content i.e., Low Fat and Regular Fat are written in distinct ways in the categorial data plots. It is observed that maximum items have low fat content.



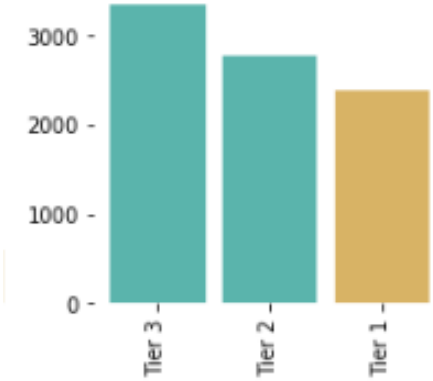
The following figure shows how each item kind is distributed. Fruits and vegetables make up the majority of the goods, followed by snack foods. Seafood, in comparison, is least in number.



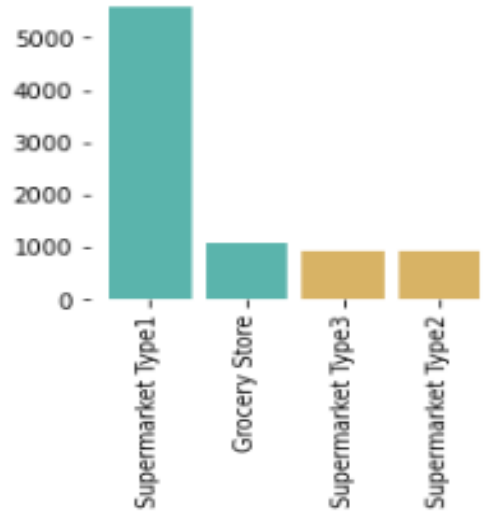
Very few of the outlets are high or large in size, whereas the majority are medium in size.



According to the statistics, there are three types of outlet locations: Tier1, Tier2, and Tier3. The Tier3 location type has the most outlets.

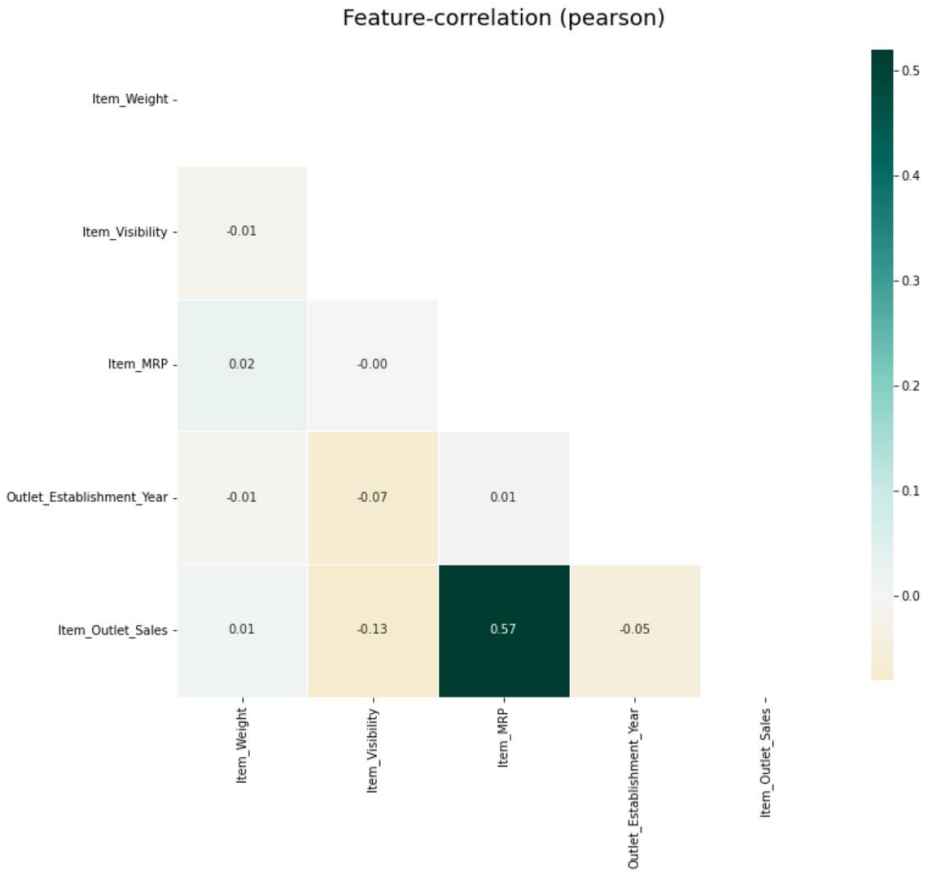


Plotted is the distribution of the several outlet types, such as Supermarket Type1, Supermarket Type2, Grocery Store, and Supermarket Type3. It has been noted that Supermarket Type 1 outlets are in majority



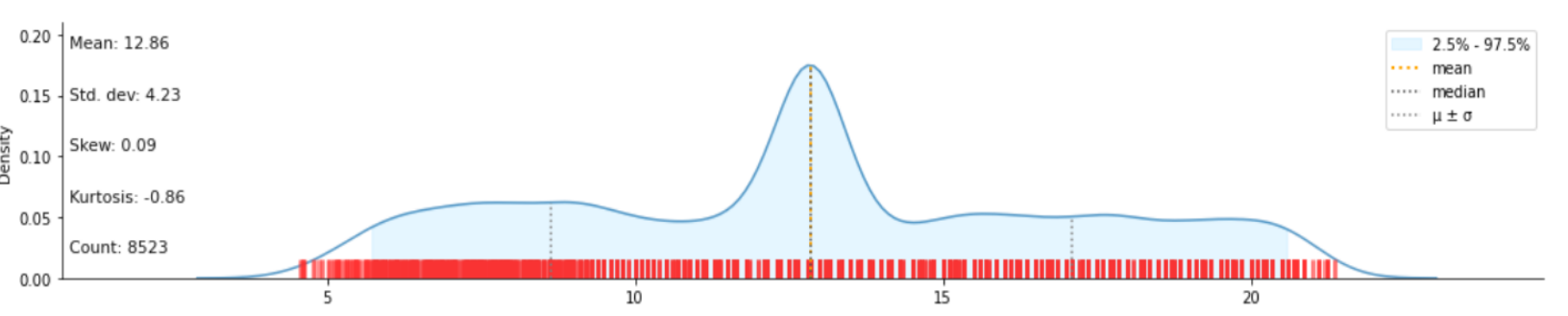
FEATURE-CORRELATION PLOTS

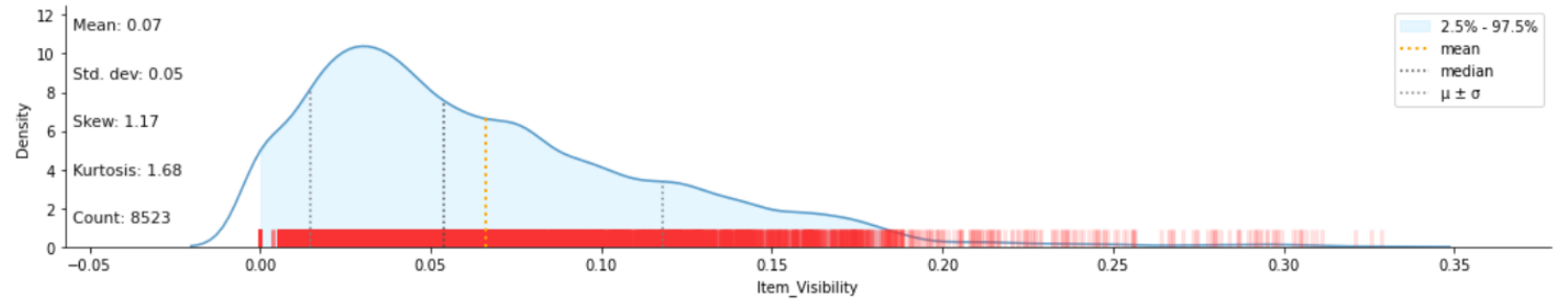
Correlation plots are used to understand which variables are connected to each other and the strength of this relationship. A correlation plot often has a number of numerical variables, with each variable represented by a column. The relationships between each pair of variables are shown by the rows. Positive values indicate a positive association, while negative values indicate a negative relationship. The values in the cells represent the strength of the relationship. You may use correlation heatmaps to identify possible links between variables and to gauge how strong these associations are.

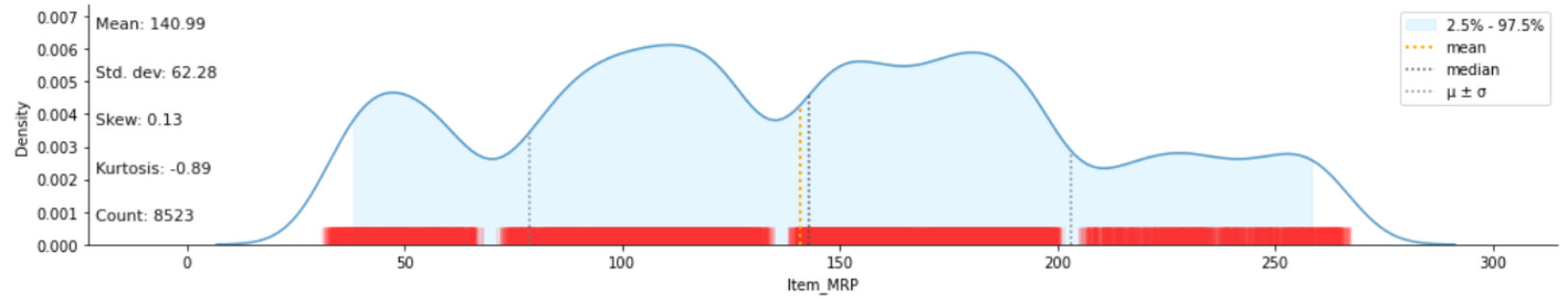


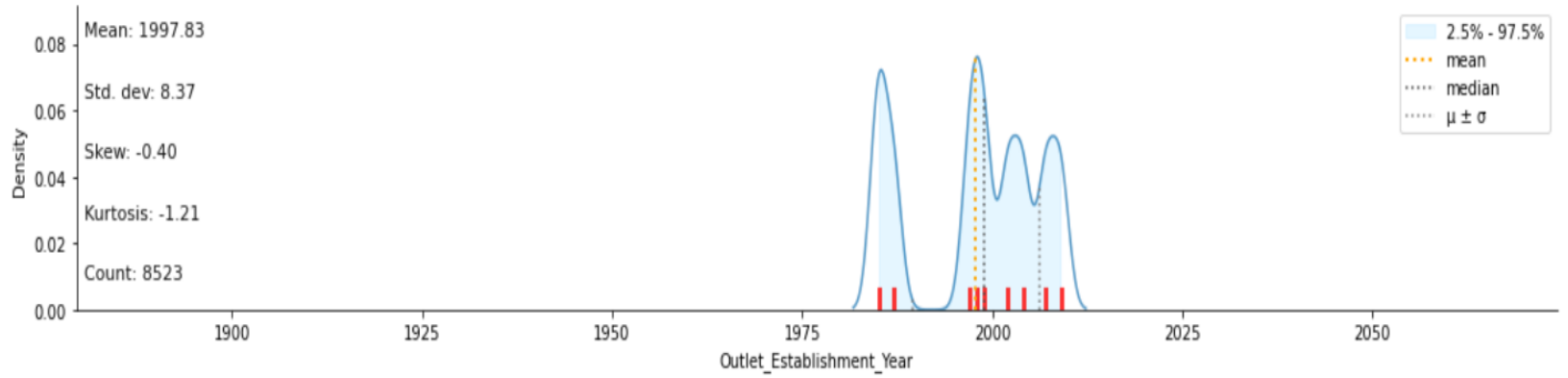
DISTRIBUTION PLOTS

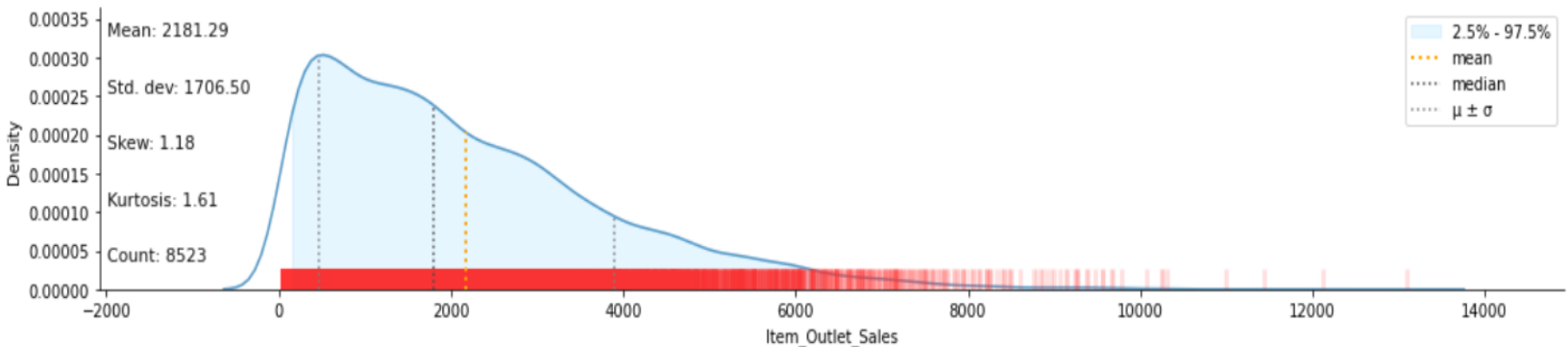
Distribution plots visually analyze the distribution of sample data by comparing the data's actual distribution to the theoretical values anticipated from a certain distribution. To ascertain if the sample data belongs to a certain distribution, use distribution plots in addition to more formal hypothesis testing. The mean, median, standard deviation, and other statistics are also provided.


Mean: 12.86
Std. dev: 4.23
Skew: 0.09
Kurtosis: -0.86
Count: 8523
2.5% - 97.5%
mean
median
μ ± σ
Density


Mean: 0.07
Std. dev: 0.05
Skew: 1.17
Kurtosis: 1.68
Count: 8523
2.5% - 97.5%
mean
median
μ ± σ
Density
Item_Visibility


Mean: 140.99
Std. dev: 62.28
Skew: 0.13
Kurtosis: -0.89
Count: 8523
2.5% - 97.5%
mean
median
μ ± σ
Density
Item_MRP


Mean: 1997.83
Std. dev: 8.37
Skew: -0.40
Kurtosis: -1.21
Count: 8523
2.5% - 97.5%
mean
median
μ ± σ
Density
Outlet_Establishment_Year


Mean: 2181.29
Std. dev: 1706.50
Skew: 1.18
Kurtosis: 1.61
Count: 8523
2.5% - 97.5%
mean
median
μ ± σ
Density
Item_Outlet_Sales

DATA PREPROCESSING

In machine learning algorithm, data can't be used in its normal form as it is the as the way it is obtained, so the data needs to be devised before employing it in machine learning models. This technique is used to solve problems that are not yet known by the knowledge extractor. This is called preprocessing work. The goal of preprocessing is to find out what kind of information the algorithm needs before making any decisions about whether to use it or not.

Preprocessing requires clean, well formatted data. The following tasks are included in data preprocessing:

1) Importing the dataset: To check the potential sales or demand of an item outlet, we used the dataset gathered from a grocery store in our study. It has the following characteristics:

   Item_identifier, Item_weight, Item_fat_content, Item_visibility, Item_type, Item_MRP, Outlet_identifier, Outlet_establishment_year, Outlet_size, Outlet_location_type, Outlet_type, Item_Outlet_sales

   The dataset file is saved as a CSV file before being imported.

2) As a part of data cleaning, it is necessary to delete some columns that don't help the algorithm reach its conclusions. Item_identifier and Outlet_identifier are removed in this case.

3) Handling missing values: Data gaps are something that must be changed to ensure that there is no disparity in the data that will be used to feed the model. It's here a few values in the fields Item_weight and Outlet_size were absent. Outlet size is an example of where the entire row has been dropped together with missing value cases and in the event of the mean of all the missing spaces for item weight is used. The other columns' entries.

4) Data Integration: One of the data preparation procedures called data integration is used to combine data from several sources into a single, bigger data storage, such as a data warehouse.

5) Data Transformation: Once the data has been cleared, we must use data transformation techniques to change the value, structure, or format of the data in order to combine the quality data into other forms.

6) Generalization: We used idea hierarchies to translate low-level or granular data to high-level information. We can transform the primitive data in the address like the city to higher-level information like the country.

7) Normalization: It is the most significant and extensively used data transformation method. Depending on the range, the numerical attributes are scaled up or down. In this method, we limit our data attribute to a certain container to create a correlation between several data points.

CONCLUSION

Demand forecasting is one of the major challenges faced by supply chains in the retail sector when trying to maximize stock levels, save costs, and boost revenue, profits, and customer loyalty. The solution to this problem is to examine and understand complex relationships and patterns from historical data using techniques like time series analysis and machine learning. In other words, it is critical to have the capability to determine what customers will buy, when they'll need it, and how much they will demand from a specific retailer or store. In order to mitigate the risks of forecasting errors, it is critical for supply chains to have a solid understanding of their end customers and how this understanding will allow them to manage uncertainty when forecasting demand from consumers. This is typically a three-part process in which data scientists first analyze and understand historical demand patterns, then use mathematical modeling to determine the impacts of various factors on the forecasted demand, and finally evaluate the accuracy of their forecasts using historical data and live data.
In this study we have used machine learning techniques to predict the sales of a particular item in a store. It will highly beneficial for increasing profits and sales of the store.

REFERENCES

Huber, Jakob, and Heiner Stuckenschmidt. "Daily retail demand forecasting using machine learning with emphasis on calendric special days." *International Journal of Forecasting* 36.4 (2020): 1420-1438.

Zhu, Xiaodan, et al. "Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry." *Production and Operations Management* 30.9 (2021): 3231-3252.

Kilimci, Zeynep Hilal, et al. "An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain." *Complexity* 2019 (2019).

Böse, Joos-Hendrik, et al. "Probabilistic demand forecasting at scale." *Proceedings of the VLDB Endowment* 10.12 (2017): 1694-1705.