University Of Chittagong

Course:CSE M398

Sales Forecasting Using Data Mining & Machine Learning Techniques

Submitted by:

Md. Nazrul Islam
ID: 14701085
Session: 2017-2018
Department of Computer Science and Engineering
University of Chittagong
Chittagong, Bangladesh

Supervisor:

Rokan Uddin Faruqui
Associate Professor
Department of Computer Science and Engineering
University Of Chittagong
Chittagong, Bangladesh

In partial fulfillment of the requirements of the degree of MS (Engineering)

in Computer Science and Engineering

Chittagong-4331,Bangladesh

Approval

This project report entitled “Sales Forecasting Using Data Mining and Machine learning Techniques” prepared and submitted by Md. Nazrul Islam, ID No: 14701085, Session 2017-2018 has been approved for submission to the Department of Computer Science and Engineering, University of Chittagong, in partial fulfillment of the requirements for the 3rd Semester MS Engineering Examination 2020.

.

Signed: 

Date: **May 11, 2022**

Rokan Uddin Faruqui
Associate Professor
Department of Computer Science and Engineering
University of Chittagong
Chittagong, Bangladesh.
Email: rokan@cu.ac.bd

CANDIDATE'S DECLARATION

The project submitted here with is a resulted of my efforts under the supervision of my superior and the help of my classmates. All information that has been obtained from other sources had been fully acknowledged. I honorably declare that this project paper is my own. I understand that any plagiarism, cheating or collusion or any sorts of constitutes a breach of university rules and regulations and would be subjected to disciplinary actions.

Md. Nazrul Islam

ACKNOWLEDGEMENTS

At the very beginning, I wish to express my deepest sense of gratitude to the Almighty Allah, for giving me great opportunity to complete this work on proper time and establishing me the chance under tremendous supervision of my honorable teacher and to give me the opportunity to work on enormous task about Sales Forecasting System using Data Mining & Machine Learning Techniques. I wish to express my supreme gratitude to my honorable supervisor Rokan Uddin Faruqui, Associate Professor, Department of Computer Science and Engineering, whose sincerity and encouragement I will never forget. I would like to thanks our honorable Professor Anwarul Azim, Chairman of Computer Science and Engineering, University of Chittagong for guiding us writing this report successfully. I would also like to thanks all of classmates for their continuous support and skilled contribution. Finally, I wish with profound respect, express my greatest gratification to my beloved parents whose moral support and blessings were a great source of inspiration during this project.

Abstract

Industrial engineering is about enhancing a process and improving the return of investment both to make more profit. A Sales Forecasting is a projection of the expected customer demand for products or services at a specific company, for a specific time horizon, and with certain underlying assumptions Essential tool used for business planning, marketing, and general management decision making. Sales Prediction can help you achieve sales goals. Prediction can help drive sales revenue, improve efficiency, increase customer retention and reduce costs. In order to predict the volume of sales, we had to determine the importance and impact of the characteristics of the products presented in the database. However, the products alone don’t explain the sales; we had to also take into consideration the contribution of the type of each outlet, its location, size and number of running years. We used Data science and statistical learning on Python software to create the predictive models that allowed us to have a better understanding of the client’s behavior and an estimation of the store’s future sales.

CONTENTS

CHAPTER: THREE

DESCRIPTION OF THE PROPOSED SYSTEM

CHAPTER: FOUR

DEVELOPMENT OF THE PROPOSED SYSTEM

CHAPTER: FIVE

EXPERIMENTAL RESULTS AND DISCUSSIONS

CHAPTER: SIX

CONCLUSION

LIST OF FIGURE

CHAPTER: ONE

INTRODUCTION

1.1 Introduction

Sales Prediction is playing a growing and important role in many fields, such as economic Prediction, electric power Prediction, resource prediction, etc. Sales prediction is an important prerequisite for enterprise planning and correct decision making, allowing companies to better plan their business activities. Prediction is important for offline businesses, especially car sales, real estate, and other everyday ventures .The forecasting are generally done by applying statistical methods, such as regression or the autoregressive–moving-average (ARMA) based on historical sales data [1]. However, these methods only work for particular data. So many factors with complex interrelationships influence sales and probably include ones with a fair degree of uncertainty. Using Machine Learning, we can identify potential models and development regularity from the masses of data. Therefore, an increasing number of researchers focus on how to make full use of Machine Learning to process historical data and handle trends in sales prediction. Data is very important for every organization and business. Data that was measured in gigabytes until recently, is now being measured in terabytes, and will soon approach the peta byte range. In order to achieve our goals, we need to fully exploit this data by extracting all the useful information from it. Unfortunately, the size and complexity of the data is such that it is impractical to manually analyze, explore, and understand the data. As a result, useful information is often overlooked, and the potential benefits of increased computational and data gathering capabilities are only partially realized. Sale data classification has different market trends. Some clusters or segments of sale may be growing, while others are declining. The information produced is very useful for business decision making. Decision can take place on the basis classification of Dead-Stock (DS), Slow- Moving(SM) and Fast-Moving (FM) of the sale [2]. Segment by-segment sales forecasting can produce very useful information. The forecasting can be short term, midterm and long term. Long term forecasting may not produce accurate predictions. However it is very useful in understanding market trends. It is easy to turn cash into inventory, but the challenge is to turn inventory into cash. Only through Machine Learning techniques, it is possible to extract useful pattern and association from the stock data. Machine Learning techniques likes clustering and associations can be used to find meaningful patterns for future predictions. Clustering is used to generate groups of related patterns, while association provides a way to get generalized rules of dependent variables. Patterns from a huge stock data on the basis of these rules can be obtained. This is a useful approach to distinguish the selling frequency of items on the basis of the known attributes, e.g. we can

examine that a “black coat of imperial company in winter season has high ratio of sale”, here we have basic property related to this example, i.e. color, type, company, season, and location. Similarly we can predict that certain products of certain properties have what type of sale trends in different locations. Thus on the basis of this scenario we can predict the reason of dead-stock, slow moving and fast moving items. Machine Learning techniques are best suited for the analysis of such type of classification, useful patterns extraction and predictions.

1.2 Challenges and Motivation

There are many challenges in the retail store network planning some of them are retailers fail in the evaluation of the potential of the market. Retailers ignore the seasonal randomness. The supply chain inefficiencies when the products have great demand then they are not available. The human resources are inefficient the employees are not available whenever necessary. The retailers face the difficulties in inventory management system; sometimes the retailers ignore the competition in the market. Retailers develop the plans that promotes the success and the highly target plan. The plans should be such that they help to obtain the maximum profit. The new product lines should be developed or they should be purchased with confidence. The supply chain mechanism should be efficient. In the retail context, an erroneous determination of the amounts to buy of each article from the suppliers, either by excess or defect can result in unnecessary costs of storage or lost sales, respectively. Both situations should be avoided by a company, which promotes the need to accurately determine the purchase quantities. Currently companies collect huge amounts of data referring to their sales and products’ features. In the past, that information was seldom analyzed and integrated in the decision making process. However, the increase of the information processing capacity has promoted the use of data analytics as a means to obtain knowledge and support decision makers in achieving better business outcomes. In addition to exploring these data and in order to perceive future trends, there is a huge volume of data available on the web that also has potential. Online content analysis can reveal trends that affect customers’ purchasing pattern and consequently this analysis is of utmost importance to support sales prediction. In this context, the development of models which use the different factors which influence sales and produce precise predictions of future sales represents a very promising strategy. The results obtained can be very valuable to the companies, as they enable them to align the amount to buy from the suppliers with the potential sales. The purpose of this study is to develop a sales

Prediction model capable of estimating the sale of new fashion products for the upcoming season.

1.3 Problem Statement

Salespeople are the headlights of every business. They have an insider's perspective of their customers and prospects. They are the first to know about what customers plan to buy in the future. Unfortunately, many companies are struggling to translate their salespeople's predictions into a reliable financial plan. World-class companies can't afford to make big forecasting errors. They create a culture of measurement in which sales and marketing manager's work in synch to achieve their goals. Higher forecasting accuracy allows them to catch market changes earlier. While many organizations are stuck tracking Prediction with spreadsheets, world-class companies use purposefully built, Sales Prediction software, such as Salesforce.com, that's integrated with their CRM solution. These solutions both capture the data and provide exceptional analytics to drive better insight and accurate action [3].

As the economy moves into the recovery phase, now is the time to invest in better sales-Prediction tools that will eliminate expensive errors and give your company the flexibility to respond quickly to market changes. Run your company based on science, not on hunches, so you can cut costs and maximize productivity. I want to find what drives the sales amount for a certain product in different stores and try to predict where and how I can maximize the sales for this particular product. The task is to predict the sales of a certain product at a particular store, part of a chain of stores and find out what influences that sale. For my research I collect some dataset from kaggle. I have access to collected data, for 1559 products across 10 stores in different cities. I will evaluate the model for the predictive accuracy using Root Mean Square Error. Assumptions: [4]

1. I'm using only grocery type products, with few features;
2. The category of the product might have an impact on sales (like dairy sells more than canned food because is used more often)
3. The type of sore and its location is important for sales

1.4 Existing System

Currently sales prediction has always used a single prediction algorithm. However, it is not possible for a single algorithm to provide sales forecasting for all kinds of commodities accurately. Thus selecting a single algorithm for a set of merchandise can be a challenge. In fact, each algorithm produces a different result, and some algorithms can produce more than one type of result. For example, a Decision Trees algorithm can not only be used for

prediction, but also as a way to reduce the number of columns in a dataset, because the decision tree can identify columns that do not affect the final mining model. So, a general prediction algorithm for all commodities is needed. Based on the strengths of each type of Machine Learning algorithms, we propose to prepare a trigger system. This trigger system analyses the data sets available for any commodity and after taking the results into account triggers the best algorithm which can be used for predicting the best accurate sales prediction for that particular commodity. The data set for the sales of commodities is freely available on many open source websites like quandl. The output of this project will be an indigenous system which will mine the data and analyze it. After analysis, it will trigger the appropriate algorithm for the sales prediction of the commodity in future time period. This will result in increased accuracy. With an increased accuracy in prediction results obtained from these algorithms, the owner can order the goods according to the predictions made and can thus avoid potential losses.

1.5 An overview of my system

I build a predictive model combining some regression models which find out the sales of each product at a particular store. Using this model, Super Shop Company will try to understand the properties of products and stores which play a key role in increasing sales. It should provide a good insight in what drives the sales for a grocery product. This is an easily scalable model to provide detailed info and accurate predictions for sales volume for different type of products as there is a lot of data out there. This solution can be used for projects, start-ups and sales prediction. For building predictive model I Replaced the Nans, identified outliers, feature selection and normalization - for both train and test data. Built the regression models: linear and decision tree. Predicted the sales, cross validated the scores, calculated the R^2 (coefficient of determination - better when using decision tree regression).Classified the train data with a decision tree and a random forest and calculated the accuracy score and the R^2. (The clear winner is the decision tree classifier)

1.6 Organization of the report

I try to organize this report in a specific way so that the reader can easily understand about the thesis and the system. The following chapters cover different aspects about the thesis. These are:

Chapter 2: Literature Review: This chapter describes the related work in this field

Chapter 3: Description of the proposed system: This chapter describes the methods used to perform the thesis.

Chapter 4: Development of the proposed system: This chapter describes how the proposed system will be developed.

Chapter 5: Experimental result and discussions: This chapter discuss about result and discussion of the research.

Chapter 6: Conclusion: This chapter discuss about conclusion and future work of the research

CHAPTER: TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter I will describe and summarize relevant work that has been done in the field of Sales Forecasting

2.2 Relevant Work

In this section, we will briefly review the previous studies on sales prediction and several classic prediction models. More than 200 kinds of prediction methods have been developed, which can be divided into two categories, subjective and objective methods. The subjective prediction method is based on the experience of experts who judge and estimate. It is strongly subjective and flexible. Examples are the Delphi method (Linstone & Turof,1975)[9], the brain storm method (Tremblay, Grosskopf, & Yang, 2010),[10] the subjective probability method (Hogarth, 1975),[5] and so on. These methods use the experience of experts or the integration of predicted results. In contrast, the objective prediction method uses raw data to build models based on mathematics and mathematical statistics methods. It is reusable but not flexible. The objective prediction method includes mainly regression analysis and time series analysis. These methods use actual sales data, establishing a reusable model in order to predict future sales. Regression analysis methods include a simple regression model, a multivariate regression model, etc. (Kleinbaum, Kupper, Nizam. Etal, 2013).[7] The time series analysis prediction model includes the moving average model, the exponential smoothing model, the seasonal trends model, the autoregressive-moving-average model the generalized autoregressive conditional heteroscedastic model, etc. (Box, Jenkins, & Reinsel, 2013)[3]Most conventional sales prediction methods introduce either factors or time series to determine the forecast. McElroy and Burmeister (1988)[6] applied Arbitrage Pricing Theory into a multivariate regression model. Lee and Fambro (1999) used the autoregressive-integrated-moving-average model to do traffic volume forecasting. In 2003, Huang and Shih (2003) forecast short-term loans using ARMA. Tay and Cao (2001)[10][8] studied time series forecasting. However, the relationship between influence factors or past time series data and sales prediction results is quite complicated. Therefore the predictions obtained from the aforementioned methods are often not satisfactory. As a consequence, many new intelligent model methods have recently been put to use in the area of forecasting; these perform better in terms of control and recognition. Some of the most representative new models are, for example, artificial neural networks (ANN) and support vector machines (SVM), the hot spots of prediction research in recent of years. Kuo and Xue (1998) [11]put forward a decision

support system for sales prediction using fuzzy neural networks. Hill, Marquez, and O'Connor (1994) reviewed the artificial neural network models for prediction and decision making. Cao (2003) [15] combined SVM with time series for sales prediction while Gao et al. (2014)[12] recommend extreme learning machine for sales prediction. Finally, Yuan (2014) proposed an online user behavior-based data mining method to predict sales in e-commerce. However, the above research focused mainly on improving the accuracy of sales prediction via optimizing a single model algorithm or analyzing the factors that influence sales. For special cases, such as when the sales volume was zero, the single prediction model didn't perform well. In addition, most of the previous methods only predicted results for one object, for example, one kind of book's sales. In actual situations, the approach needs to cover a large scale of products. Thus, the traditional single model optimization method has significant limitations in sales prediction. I built a predictive model system instead of depending on a single model algorithm. Based on data about factors that influence sales, "the system" triggers one of the prediction models discussed previously, leading to better prediction results than before. Also, our method can be used for a much larger scale of sales prediction. Therefore, we provide a new proposal for sales prediction research, which has been proven to be a significant improvement over past methods through our validation.

2.3 Types of Forecasts

- **Economic prediction**: Predict a variety of economic indicators, like money supply, inflation rates, interest rates, etc.
- **Technological forecasts**: Predict rates of technological progress and innovation
- **Demand prediction**: Predict the future demand for a company's products or service

2.4 Various Techniques for Sales Forecasting

There are two types of Sales Prediction Techniques one is Qualitative Techniques another one is Quantitative Techniques. Now we discuss those techniques elaborately

2.4.1 Qualitative Techniques

Qualitative Prediction techniques [13] are sometimes referred to as judgmental of subjective techniques because they rely more upon opinion and less upon mathematics in their formulations. The absence of past sales means that you have to be more creative in coming up with prediction in the future. Sales forecasting for new products are often based on

executive judgments, sales force projection, surveys and user's expectation. We summarized qualitative forecasting techniques which include:

A) Executive Opinion Method

- Most widely used
- Method of combining and averaging views of several executives regarding a specific decision or prediction.

B) Delphi Method

Process includes a coordinator getting prediction separately from experts, summarizing the prediction giving the summary report to experts who are asked to make another prediction; the process is repeated till some consensus is reached

C) Sales Force Composite Method

- Also known as "Grassroots Approach"
- Individual salespersons prediction sales for their territories

D) Survey of Buyer's Intentions

- Process includes asking customers about their intentions to buy the company's product and services
- Questionnaire may contain other relevant questions

2.4.2 Quantitative Techniques

Quantitative techniques are sometimes termed objective or mathematical techniques as they rely more upon mathematics as less upon judgment in their computation. These techniques are now very popular as a result of sophisticated computer packages. We summarized qualitative forecasting techniques which include[13]:

A) Time Series Analysis

Time series models look at past patterns of data and attempt to predict the future based upon the underlying patterns contained within those data.

B) Market Test Method

- Used for developing one time prediction particularly relating to new products
- A market test provides data about consumers' actual purchases and responsiveness to the various elements of the marketing mix.

C) Regression Analysis

- Identifies a statistical relationship between sales(dependent variable) and one or more influencing factors, which are termed the independent variables.
- When just one independent variable is considered (eg. population growth), it is called a linear regression, and the results can be shown as a line graph predicting future values of sales based on changes in the independent variable.
- When more than one independent variable is considered, it is called a multiple regression

❖ The Advantages of Regression Analysis & Forecasting

Managers need information to evaluate what is going on in the external and the internal environments of an organization. Regression analysis is one of the quantitative models that managers use to study the behavior of semi-variable costs and separate the fixed and the variable elements. Managers prefer the regression analysis technique to other models such as the high-low and scatter graph methods because of the overall superiority of the results.

1. Accuracy of Results

Regression analysis allows managers to establish objective measures of relationships between the independent and the dependent variables, rather than purely using personal judgment. This generally results in accurate information that is more reliable for decision-making, and other parties can empirically test the results using the same or separate data without resulting to personal opinions.

2. Assessment Tools

When the management obtains the results of the regression models electronically, most of the computers they use have software packages that provide a few statistics, such as the R-square and the student t-value statistics. The two statistics help managers determine the accuracy of

the predictions, and thus the level of reliability of the results that they have obtained using the regression equations.

3. Use of Multi-Variables

The multiple regression analysis models allow managers to test for several independent variables that may explain different things about the dependent variable. Though complex, the manager can test for all the factors that he thinks have an effect on a given depended variable. This is unlike other inferior models that allow for only one independent variable. With the use of several variables, the accuracy of prediction is also improved.

4. Input for New Management Trends

Regression analysis provides needed input for activity-based cost and management techniques. These techniques are based on knowing what activities or transactions cause the acquisition and use of resources. The theory of constraints encourages managers to look at throughput per scarce resource as part of dealing with a dynamic environment of changing constraints. Regression analysis allows managers to establish objective.

D) Moving Averages

takes an average of a specified number of past observations to make a forecast. As new observations become available, they are used in the forecast and the oldest observations are dropped.

E) Exponential Smoothing

- Similar to moving average method
- Used for short run prediction
- Instead of weighing all observations equally in generating the prediction, exponential smoothing weighs the most recent observations heaviest Next year's sale=a(this year's sale) + (1-a)(this year's forecast)a is smoothing constant taken in scale 0-1

.

CHAPTER: THREE

DESCRIPTION OF THE PROPOSED SYSTEM

3.1 Introduction

The proposed methodology aims to make a predictive model for Future Sales Prediction system

3.2 System Architecture

The system is divided into the following steps:

1. Making some hypothesis about what factors could potentially affect the outcome of the problem statement.

2. Collecting Data

3. Data Exploration

4. Data Cleaning

5. Feature Engineering

6. Building a predictive Model

7. Forecasting

8. Evolution of Predicted Outcomes

Here is a simple graphical representation of our system given below, which helps us to understand work flow of our system.

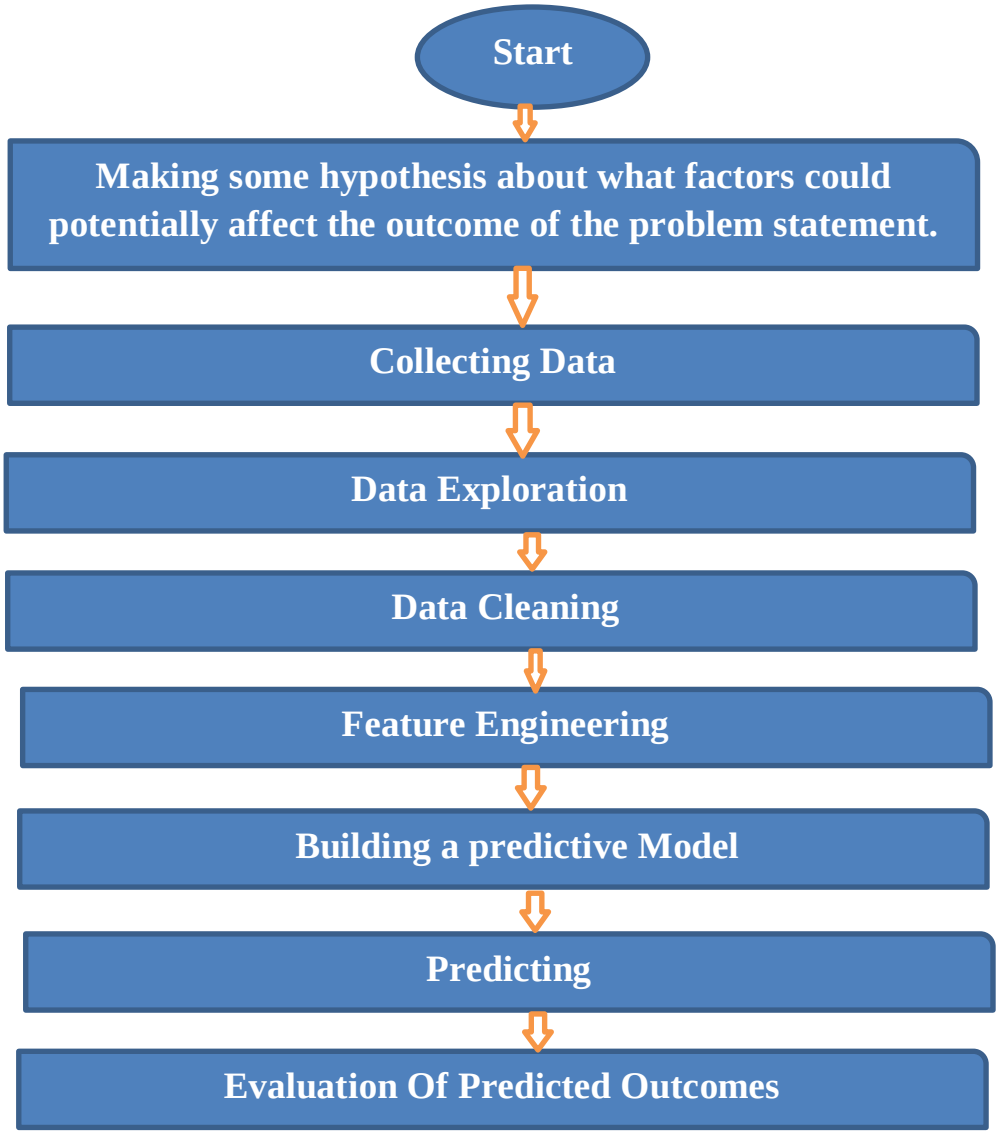


Fig: System Work flow

3.3 Hypothesis Generation

This is a crucial step in the Machine Learning [14] process. It involves understanding the problem and making some hypothesis about what factors could potentially affect the outcome of the problem statement. This step should be done before looking at the data.

❖ Store level Hypotheses:

1. **City type** Store located in Tier 1 cities should have higher sales due to higher average income level

2. **Population density** Stores that are located in high population density area should experience higher demand, therefore, higher sales

3. **Store capacity** Big stores act like a one-stop-shops and people prefer to get everything in one place, therefore big stores tend to have higher sales

4. **Competitors Stores** located in areas with similar establishments nearby should have less sales due to competition

5. **Marketing** Stores that has a good marketing team should have higher sales and it can attracts customers with the right offers and advertising

6. **Customer behavior** Stores keeping highly demanded products by local customers should experience higher sales

7. **Ambiance** Stores that are well-maintained and has great customer service are expected to have higher sales

❖ Product level Hypotheses:

1. **Brand** Popular brand products should have higher sales due to loyal and higher trust from customers

2. **Packaging** Good packaging can attract customers and sell more

3. **Utility** Daily use products should have a higher tendency to sell

4. **Display Area** Products with bigger shelves space are likely to catch customers attention and sell more

5. **Visibility in Store** The location of product in a store will impact sales (e.g. right at entrance or near the tills)

6. **Advertising** Better advertising of products should lead to higher sales

7. **Promotional Offers** that are attractive to customers will sell more

3.4 Data Collection

Business forecasting by its very nature uses historical data to forecast future performance of the company. Historical data includes your company's financial statements, client invoices and any information you believe has relative predictive value to the future success of your company. Historical data doesn't have to solely come from your company; it can also be historical macroeconomic data, such as the Consumer Confidence Index, interest rates, housing starts or any other economic variable you believe has an effect on your business based on your observations and business experience.

3.5 Data Exploration

We'll be performing some basic data exploration here and come up with some inferences about the data. We'll try to figure out some irregularities and address them in the next section. If you are new to this domain, please refer our Guide. The first step is to look at the data and try to identify the information which we hypothesized vs. the available data. In this step we describe our dataset in various ways for understanding dataset and finding missing values

3.6 Data Cleaning

This step typically involves imputing missing values and treating outliers. Though outlier removal is very important in regression techniques, advanced tree based algorithms are impervious to outliers. So I'll leave it to you to try it out. We'll focus on the imputation step here, which is a very important step.

3.7 Feature Engineering

In this section, we will make our data ready for analysis by modifying/creating new variables we explored some nuances in the data in the data exploration section. Let's move on to

resolving them and making our data ready for analysis. We will also create some new variables using the existing ones in this section.

3.8 Model Building

We built several kind of regression model in python to predict sales using outlet identifier, item identifier and price of the items. We have used backward selection model to analyze the effects of various predictors on the sales. Regression analysis applies to almost any field. In general, regression analysis identifies relationships based on independent and dependent variables. For example, a dependent variable is your company's sales and an independent variable may be interest rates. Governments and businesses use regression analysis as a predictive tool for forecasting purposes. Regression analysis measures the strength or correlation between the dependent and independent variables. For the uninitiated, regression analysis may become complex particularly with the addition of multiple independent and dependent variables, requiring sophisticated programs and analysis.

3.9 Forecasting

Business forecasting using historical data is tricky business. Past performance is not necessarily a good indicator of future performance. In other words, you shouldn't forecast more than 10 percent in sales growth in the next period because your company's sales grew more than 10 percent in three prior periods. This is where regression analysis comes in. By identifying the relationship between the dependent variable or your company's sales with independent variables, you may gain a better insight as to the factors that determine your company's sales growth. Generally speaking, the predictive power of regression analysis improves the longer the time period used to build your forecast.

3.10 Evaluation of Forecasting Outcomes

Using historical data and regression analysis has its limitations in business forecasting. For example, a significant correlation between the independent and dependent variable does not necessarily indicate a cause and effect relationship. In some cases, the common linkage may be because of a sequence of events. If you are considering whether to incorporate regression analysis as a forecasting tool for your business, there are several user-friendly statistical programs available to help you perform regression analysis, which includes identifying statistical errors that may affect your forecast.

CHAPTER: FOUR

DEVELOPMENT OF THE PROPOSED SYSTEM

4.1 Introduction

In this chapter, we will discuss our system development process .Actually we mainly worked on one of the Quantitative Techniques names Regression.

4.2 Tools and Materials

We will use Anaconda a version for Windows to implement our Sales Forecasting system. Especially we will use Jupyter lab/Google Colab which is responsive and faster and user friendly.

4.2.1 ANACONDA

It is a FREE enterprise-ready Python distribution for data analytics, processing, and scientific computing. Anaconda comes with Python 2.7 or Python 3.4 and 100+ cross-platform tested and optimized Python packages. All of the usual Python ecosystem tools work with Anaconda. Additionally, Anaconda can create custom environments that mix and match different Python versions (2.6, 2.7, 3.3 or 3.4) and other packages into isolated environments and easily switch between them using conda, our innovative multi-platform package manager for Python and other languages

4.2.2 PYTHON

Python can be easy to pick up whether you're a first time programmer or you're experienced with other languages. The following pages are a useful first step to get on your way writing programs with Python!

4.3 Discussion about our Dataset

In this section we will try to introduce our collecting dataset which we use in our system for building a predictive model.

4.3.1 Training Dataset Sample

Different datasets tend to expose new issues and challenges, and it is interesting and instructive to have a variety of problems when considering learning method .Here is screenshot given below from our huge dataset.

G1 fx Outlet_Identifier

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Item_Iden	Item_Weig	Item_Fat_	Item_Visib	Item_Type	Item_MRP	Outlet_Ide	Outlet_Est	Outlet_Siz	Outlet_Loc	Outlet_Typ	Item_Outlet_Sales	
2	FDA15	9.3	Low Fat	0.01605	Dairy	249.809	OUT049	1999	Medium	Tier 1	Supermark	3735.14	
3	DRC01	5.92	Regular	0.01928	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermark	443.423	
4	FDN15	17.5	Low Fat	0.01676	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermark	2097.27	
5	FDX07	19.2	Regular	0	Fruits and	182.095	OUT010	1998		Tier 3	Grocery St	732.38	
6	NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermark	994.705	
7	FDP36	10.395	Regular	0	Baking Goo	51.4008	OUT018	2009	Medium	Tier 3	Supermark	556.609	
8	FDO10	13.65	Regular	0.01274	Snack Foo	57.6588	OUT013	1987	High	Tier 3	Supermark	343.553	
9	FDP10		Low Fat	0.12747	Snack Foo	107.762	OUT027	1985	Medium	Tier 3	Supermark	4022.76	
10	FDH17	16.2	Regular	0.01669	Frozen Foo	96.9726	OUT045	2002		Tier 2	Supermark	1076.6	
11	FDU28	19.2	Regular	0.09445	Frozen Foo	187.821	OUT017	2007		Tier 2	Supermark	4710.54	
12	FDY07	11.8	Low Fat	0	Fruits and	45.5402	OUT049	1999	Medium	Tier 1	Supermark	1516.03	
13	FDA03	18.5	Regular	0.04546	Dairy	144.11	OUT046	1997	Small	Tier 1	Supermark	2187.15	
14	FDX32	15.1	Regular	0.10001	Fruits and	145.479	OUT049	1999	Medium	Tier 1	Supermark	1589.26	
15	FDS46	17.6	Regular	0.04726	Snack Foo	119.678	OUT046	1997	Small	Tier 1	Supermark	2145.21	

Fig: Training Dataset Sample

4.3.2 Testing Dataset Sample

A1 fx Item_Identifier

	A	B	C	D	E	F	G	H	I	J	K	L
1	Item_Iden	Item_Weig	Item_Fat_	Item_Visib	Item_Type	Item_MRP	Outlet_Ide	Outlet_Est	Outlet_Siz	Outlet_Loc	Outlet_Type	
2	FDW58	20.75	Low Fat	0.00756	Snack Foo	107.862	OUT049	1999	Medium	Tier 1	Supermarket Type1	
3	FDW14	8.3	reg	0.03843	Dairy	87.3198	OUT017	2007		Tier 2	Supermarket Type1	
4	NCN55	14.6	Low Fat	0.09957	Others	241.754	OUT010	1998		Tier 3	Grocery Store	
5	FDQ58	7.315	Low Fat	0.01539	Snack Foo	155.034	OUT017	2007		Tier 2	Supermarket Type1	
6	FDY38		Regular	0.1186	Dairy	234.23	OUT027	1985	Medium	Tier 3	Supermarket Type3	
7	FDH56	9.8	Regular	0.06382	Fruits and	117.149	OUT046	1997	Small	Tier 1	Supermarket Type1	
8	FDL48	19.35	Regular	0.0826	Baking Goo	50.1034	OUT018	2009	Medium	Tier 3	Supermarket Type2	
9	FDC48		Low Fat	0.01578	Baking Goo	81.0592	OUT027	1985	Medium	Tier 3	Supermarket Type3	
10	FDN33	6.305	Regular	0.12337	Snack Foo	95.7436	OUT045	2002		Tier 2	Supermarket Type1	
11	FDA36	5.985	Low Fat	0.0057	Baking Goo	186.892	OUT017	2007		Tier 2	Supermarket Type1	
12	FDT44	16.6	Low Fat	0.10357	Fruits and	118.347	OUT017	2007		Tier 2	Supermarket Type1	
13	FDQ56	6.59	Low Fat	0.10581	Fruits and	85.3908	OUT045	2002		Tier 2	Supermarket Type1	
14	NCC54		Low Fat	0.17108	Health and	240.42	OUT019	1985	Small	Tier 1	Grocery Store	
15	FDU11	4.785	Low Fat	0.09274	Breads	122.31	OUT049	1999	Medium	Tier 1	Supermarket Type1	

Fig: Testing Dataset Sample

The first step is to look at the data and try to identify the information which we hypothesized vs. the available data. A comparison between the data dictionary on the competition page and out hypotheses is shown below:

Variable	Description	Relation to Hypothesis
Item_Identifier	Unique product ID	ID Variable
Item_Weight	Weight of product	Not considered in hypothesis
Item_Fat_Content	Whether the product is low fat or not	Linked to 'Utility' hypothesis. Low fat items are generally used more than others
Item_Visibility	The % of total display area of all products in a store allocated to the particular product	Linked to 'Display Area' hypothesis.
Item_Type	The category to which the product belongs	More inferences about 'Utility' can be derived from this.
Item_MRP	Maximum Retail Price (list price) of the product	Not considered in hypothesis
Outlet_Identifier	Unique store ID	ID Variable
Outlet_Establishment_Year	The year in which store was established	Not considered in hypothesis
Outlet_Size	The size of the store in terms of ground area covered	Linked to 'Store Capacity' hypothesis
Outlet_Location_Type	The type of city in which the store is located	Linked to 'City Type' hypothesis.
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket	Linked to 'Store Capacity' hypothesis again.
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.	Outcome variable

We can summarize the findings as:

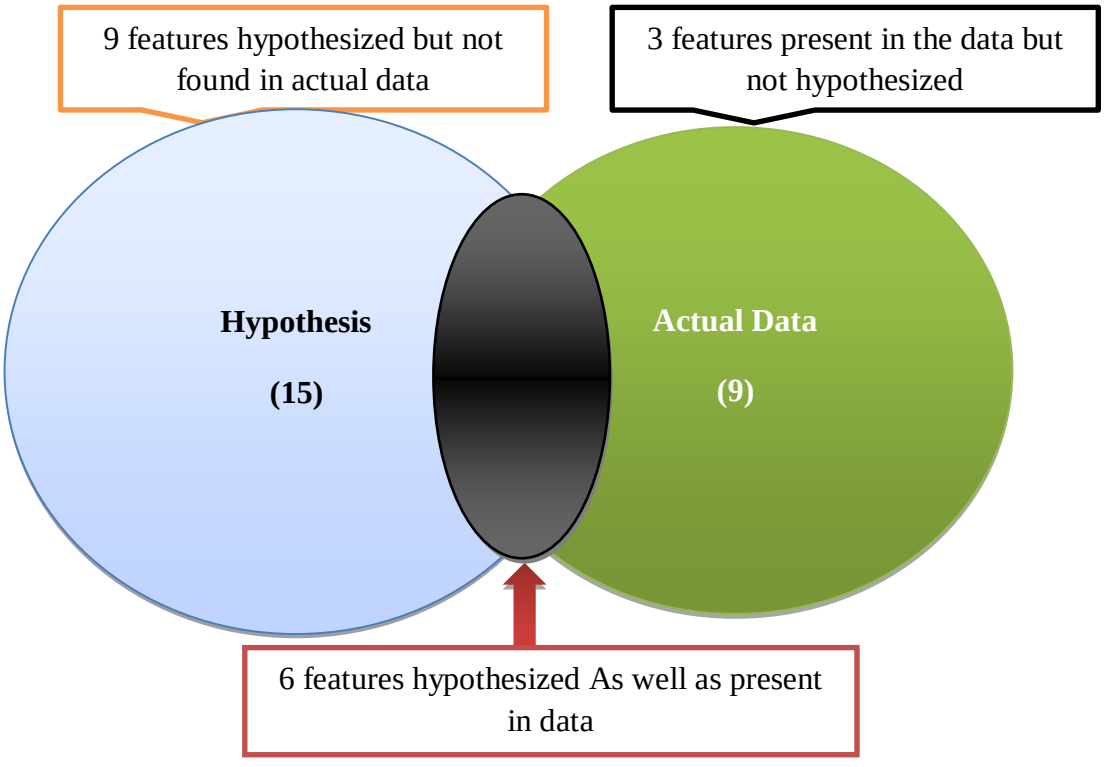


Fig: Hypothesized Vs. Data Set summery

We will invariable find features which we hypothesized, but data doesn't carry and vice versa. We should look for open source data to fill the gaps if possible.

4.4 Data Preprocessing Steps

In this section we will try to describe our data preprocessing steps for building predictive model.

4.4.1 Data Exploration

After importing some necessary python packages now, we load our training data set into our system then we print how many numbers rows and columns into training dataset

```
Sales_train = pd.read_csv('/home/mine/Desktop/Train1.csv')
Sales_train["Type"] = "train"
print(Sales_train.shape)
```

(8523, 13)It's indicates that in our training dataset there are 8523 rows and 13 columns Loading Testing dataset and find out how many numbers of rows and columns there

```
Sales_test = pd.read_csv('/home/mine/Desktop/Test1.csv')
Sales_test["Type"] = "test"
print(Sales_test.shape)
```

(5681, 12) It's indicates that in our testing dataset there are 5681 rows and 12 columns.

Now combining training dataset and testing dataset into one data frames

```
dframes = [Sales_train,Sales_test]
Sales = pd.concat(dframes,ignore_index = True)
print(Sales.shape)
```

(14204, 14) It's indicates that in our combining data frame there are 14204 rows and 14 columns.

2. Checking for missing values in Dataset

```
In [36]: Sales.isnull().sum()

Out[36]: Item_Fat_Content               0
         Item_Identifier                0
         Item_MRP                       0
         Item_Outlet_Sales           5681
         Item_Type                      0
         Item_Visibility                0
         Item_Weight                 2439
         Outlet_Establishment_Year      0
         Outlet_Identifier              0
         Outlet_Location_Type           0
         Outlet_Size                 4016
         Outlet_Type                    0
         num_years                      0
         source                         0
         dtype: int64
```

Note that the 5681 missing values Outcome variable comes from the test dataset, which is normal as those are the values we are trying to predict. In terms of Item_Weight and Outlet_Size we will impute the missing values in the data cleaning section. Below is a more visualize way of finding the missing values

3. Missing Data Visualization:

Now we visualize our missing data using python seabron package.

Run this code into our system we get missing data visualization graphical output

```
sns.heatmap(Sales.isnull().cbar=False,cmap='viridis',yticklabels=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a0aa629e8>
```

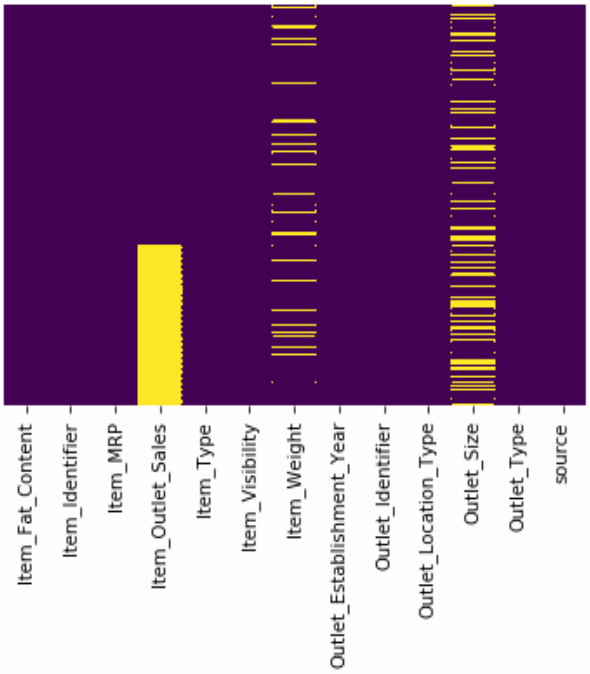


Fig: Graphical Representation of Missing Values

4. Cleaning dataset values Visualization

This step involves imputing missing values and treating outliers. Treating outliers are important for regression techniques although advanced tree based algorithms are impervious to them.

After cleaning dataset .This below diagram shows us dataset values present state. Here we see that Item_Weight and Outlet_Size columns missing values now fix. Here we also see that Item_Outlet_Sales columns still have some missing values .But this missing comes from test data set which don't need to consider.

```
sns.heatmap(Sales.isnull().cbar=False,yticklabels=False,cmap='viridis')
```

<matplotlib.axes._subplots.AxesSubplot at 0x1a0aa629e8>

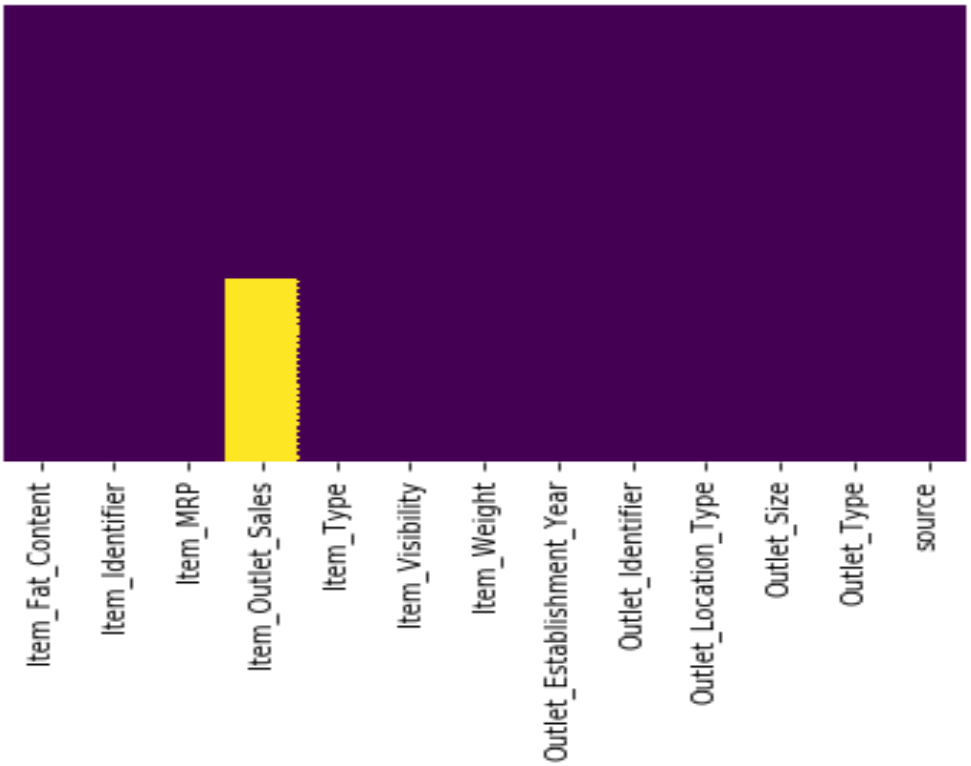


Fig: Graphical Representations of Missing values after cleaning

4.4.2 Feature Engineering Steps

In this section, we will make our data ready for analysis by modifying/creating new variables

1. Combining Outlet_Type

In the data exploration section, we decided to consider combining the Supermarket Type2 and Type3 variables. In order to check if this is a good idea we can analyze the mean sales by the type of store. If they have similar sales, then keeping them separate won't help much

2. Modify Item_Visibility

In our dataset We noticed at the beginning that the minimum value for Item_Visibility is 0, which doesn't make any sense. We have decided to treat the 0 like missing information and impute it with mean visibility of that product.

3. Years of store operation

The latest year within our data is 2018 so we can use this and the establishment year variable to calculate the years of operation of a store. The result shows that store in our dataset are 9 – 31 years old.

Sample code for Outlet store present age counting.

```
In [230]: import datetime
          now = datetime.datetime.now()
          now.year

          Sales["Outlet_Age"] = now.year - Sales["Outlet_Establishment_Year"]
          Sales["Outlet_Age"].head()

Out[230]: 0    19
          1     9
          2    19
          3    20
          4    31
          Name: Outlet_Age, dtype: int64
```

4. Modify Item_Fat_Content

Earlier, we spotted that there are some mislabeling in the **Item_Fat_Content** variable. In addition, we noticed that some non-consumables are labeled as 'Low Fat' content which doesn't make sense so we are going to fix this too by creating a separate category to spot this!

Original Categories:

```
Low Fat    8485
Regular    4824
LF          522
reg         195
low fat     178
Name: Item_Fat_Content, dtype: int64
```

Modified Categories:

```
Low Fat    9185
Regular    5019
Name: Item_Fat_Content, dtype: int64
```

5. Numerical and One-Hot Coding of Categorical variables

Since scikit-learn accepts only numerical variables, I converted all categories of nominal variables into numeric types. One-Hot-Coding refers to creating dummy variables, one for each category of a categorical variable. For example, the Item_Fat_Content has 3 categories – 'Low Fat', 'Regular' and 'Non-Edible'. One hot coding will remove this variable and generate 3 new variables. Each will have binary numbers – 0 (if the category is not present) and 1(if category is present). This can be done using 'get_dummies' function of Pandas. Here is Screen shots given of this process.

```
In [116]: Sales[['Item_Fat_Content_0','Item_Fat_Content_1','Item_Fat_Content_2']].head()
```

Out[116]:

	Item_Fat_Content_0	Item_Fat_Content_1	Item_Fat_Content_2
0	1	0	0
1	0	0	1
2	1	0	0
3	0	0	1
4	0	1	0

We can notice that each row will have only one of the columns as 1 corresponding to the category in the original variable

6. Checking Data Types after Preprocessing

```
In [115]: Sales.dtypes

Out[115]: Item_Identifier              object
          Item_MRP                    float64
          Item_Outlet_Sales           float64
          Item_Type                    object
          Item_Visibility             float64
          Item_Weight                 float64
          Outlet_Establishment_Year     int64
          Outlet_Identifier            object
          num_years                     int64
          source                       object
          Years_of_Operation            int64
          Item_Fat_Content_0            uint8
          Item_Fat_Content_1            uint8
          Item_Fat_Content_2            uint8
          Outlet_Location_Type_0        uint8
          Outlet_Location_Type_1        uint8
          Outlet_Location_Type_2        uint8
          Outlet_Size_0                 uint8
          Outlet_Size_1                 uint8
          Outlet_Size_2                 uint8
          Item_Type_Category_0          uint8
          Item_Type_Category_1          uint8
          Item_Type_Category_2          uint8
          Outlet_Type_0                 uint8
          Outlet_Type_1                 uint8
          Outlet_Type_2                 uint8
          Outlet_Type_3                 uint8
          Outlet_0                      uint8
          Outlet_1                      uint8
          Outlet_2                      uint8
          Outlet_3                      uint8
          Outlet_4                      uint8
          Outlet_5                      uint8
          Outlet_6                      uint8
          Outlet_7                      uint8
          Outlet_8                      uint8
          Outlet_9                      uint8
          dtype: object
```

7. Dataset variables Correlation Visualization

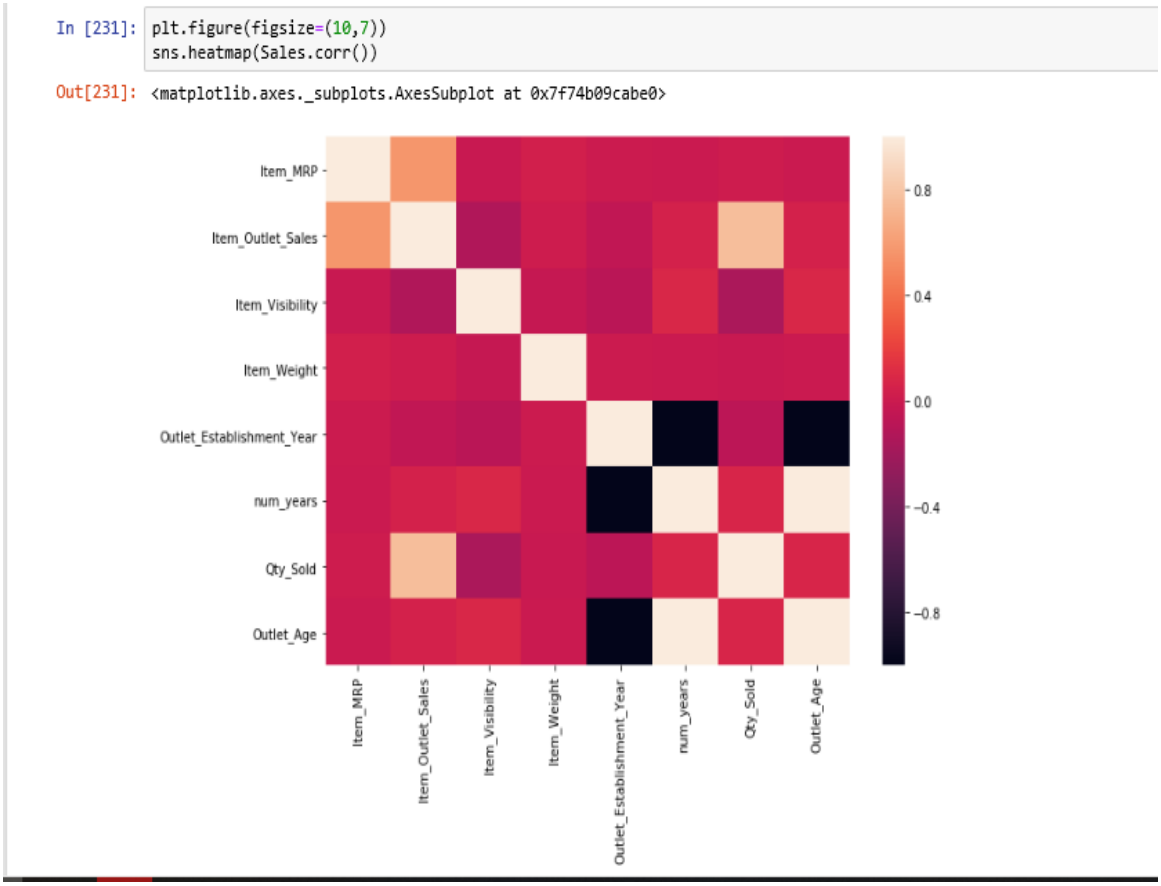


Fig: Dataset variables correlation heatmap

4.5 Exporting Data

Final step of this section is to convert the data back to train and test datasets. We also need to do some final tidying of deleting some columns before and after the split.

4.6 Data Classification and Model Building

Here's where we build our predictive model and find out accuracy of our dataset. We will be going through 5 models which include linear regression, decision tree, Lasso, Ridge, and random forest

4.6.1 KNN Cross Validations Process

1. Read the training data from a file

2. Read the testing data from a file

3. Set K to some value set the learning rate α

4. Set the value of N for number of folds in the cross validation

5. Normalize the attribute values in the range 0 to 1 n Value = Value / (1+Value)

6. Calculate the accuracy as n Accuracy = (# of correctly classified examples / # of training examples) X 100

7. Repeat the process till desired accuracy is reached

4.6.2 Simple models for Prediction

Let us start with making predictions using a few simple ways to start with. If I were to ask you, what could be the simplest way to predict the sales of an item, what would you say?

1. Model 1 – Mean sales:

Even without any knowledge of machine learning,[15] you can say that if you have to predict sales for an item – it would be the average over last few days. / Months / weeks. It is a good thought to start, but it also raises a question – how good is that model? Turns out that there are various ways in which we can evaluate how good our model is. The most common way is Mean Squared Error. Let us understand how to measure it.

❖ Prediction error

To evaluate how good a model is, let us understand the impact of wrong predictions. If we predict sales to be higher than what they might be, the store will spend a lot of money making unnecessary arrangement which would lead to excess inventory. On the other side if I predict it too low, I will lose out on sales opportunity. So, the simplest way of calculating error will be, to calculate the difference in the predicted and actual values. However, if we simply add

them, they might cancel out, so we square these errors before adding. We also divide them by the number of data points to calculate a mean error since it should not be dependent on number of data points.[15]

$$\frac{(e_1^2 + e_2^2 + \ldots + e_n^2)}{n}$$

[Each error squared and divided by number of data points]This is known as the mean squared error. Here e1, e2 …. , en are the difference between the actual and the predicted values. So, in our first model what would be the mean squared error? On predicting the mean for all the data points, we get a mean squared error = 29, 11,799. Looks like huge error. May be its not so cool to simply predict the average value. Let's see if we can think of something to reduce the error.

2. Model 2 – Average Sales by Location:

We know that location plays a vital role in the sales of an item.[16] For example, let us say, sales of car would be much higher in Delhi than its sales in Varanasi. Therefore let us use the data of the column 'Outlet_Location_Type'.So basically, let us calculate the average sales for each location type and predict accordingly. On predicting the same, we get mse = 28,75,386, which is less than our previous case. So we can notice that by using a characteristic [location], we have reduced the error. Now, what if there are multiple features on which the sales would depend on. How would we predict sales using this information? Regression models come to our rescue.

4.6.3 Regression and Correlation Analysis

The term regression analysis is used to refer to studies of relations between variables. Regression analysis then is a technique for quantifying the relationship between a criterion variable (also called a dependent variable) and one or more predictor variables (or independent variables). Put another way regression analysis is a technique whereby a mathematical equation is fitted to a set of data. It describes the relationship between two variables. It is a statistical method used to isolate cause- and –effect relation among variables. A line of best fit that is independent of individual judgment and drawn mathematically is called a regression line. Regression line equations once computed can be graphed and be used to estimate values previously unknown[17].

The reasons for computing a regression line are:

(i) To obtain a line of best fit free of subjective judgment. The regression line improves our estimates.

(ii) The regression equation can be used to make predictions within the given range of the data that is, making interpolations.

(iii) The reliability of estimates made from such a line can be measured mathematically.

The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the available data. Values of Y will be predicted from the values of X, hence the closer the points are to the line, the better the fit and the predictions will be.

Each observation of bivariate data can be viewed as a point (x, y) where x is the explanatory or independent variable while y is the dependent or response variable. It is important to determine which one of the variables in question is the dependent variable and which one is the independent variable. The starting point in regression analysis is the construction of a scatter diagram/ scatter graph.[18]

There are normally two regression line equations for any sets of bivariate data[18]:

(1) Regression of Y on X. The line equation is given by $Y= a+ bX$ where the line is used to predict/ estimate the value of Y that follows from any value of X

(2) Regression of X on Y. The equation is given by $X = a + bY$. The line equation is used to predict/ estimate the value of X that follows from any value of Y.

The guide on the choice of regression line equation is one should always use the line that has the variable to be estimated on the left hand side of the equation. In other words if you want to estimate X use $X = a + bY$ and if you want to estimate Y use $Y = a + bX$. In our analysis we are going to use the Y on X regression equation, that is $Y = a + bX$. In this equation a is the y intercept term, while b is called the regression coefficient or the slope of the graph.

In single equation regression models one variable called the dependent variable or regressand is expressed as a linear function of one or more other variables called the explanatory variables or independent or regressor variables. It is assumed implicitly that causal relationships if any between the independent and dependent variables flow in one direction only, namely from the explanatory to the dependent (or regressand).

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. A statistical relationship may be very strong but it never can establish causal connection. Our ideas of causation must come from outside statistics, ultimately from some theory or other. For example there is no statistical reason to assume that rainfall depends on crop yield. In regression analysis we try to estimate or predict the average value of one variable on the basis of fixed values of other variables. The regression equation is given by Y = a + bX or

Y = f(X) or Yi = a + bX +μi where μi is the stochastic error term or random error term (it can also be looked at as the surrogate for all omitted variables. ß1 and ß2 are regression coefficients. ß1 is the intercept and ß2 is the slope coefficient.

E(Y|Xi) = ß1 + ß2Xi .= $\hat{Y}$

μi = Yi – E (Y|Xi), where Yi is the observed value and E (Y|Xi), is the predicted value.

Regression analysis techniques assume that

(1) Each item of data is independent of the others.

(2) Data measurements are unbiased.

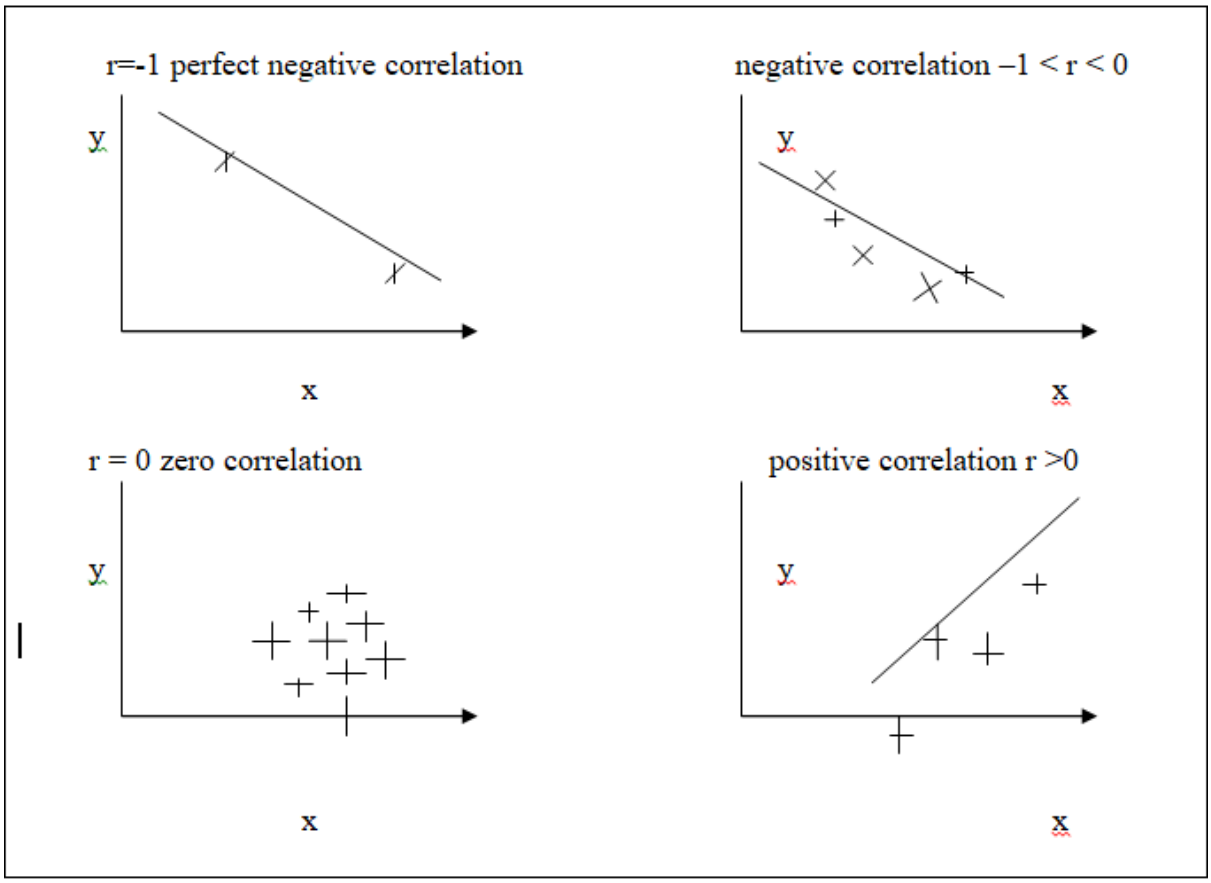
(3) The error variance is constant over the entire range of data, rather than larger in some parts of the data range and smaller in others.

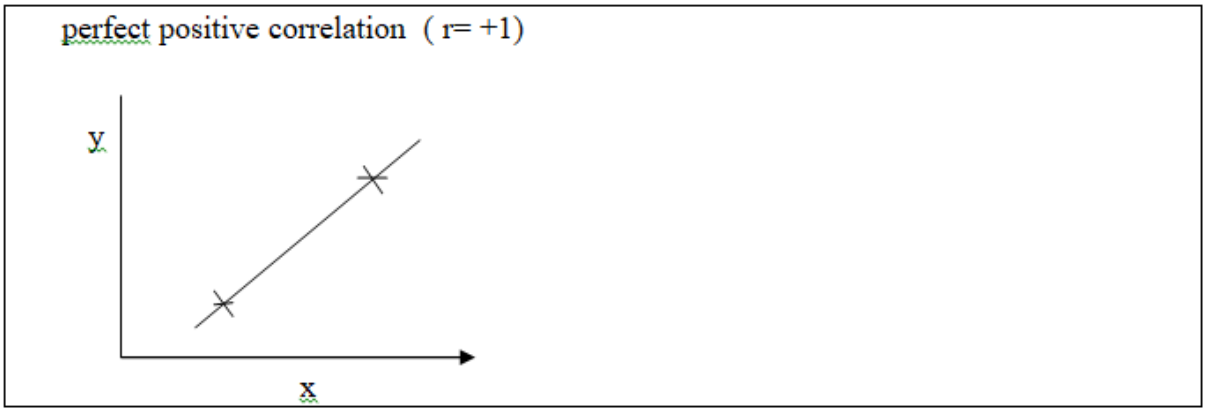
(4) There is no autocorrelation between the disturbances (or errors).

❖ **Correlation Analysis**

Correlation is a statistical method used to determine whether a relationship between variables exists.[18][19] Correlation is designed to measure the strength or degree of linear association between two variables. While regression analysis establishes a mathematical relationship between the dependent and independent variables, correlation goes a step further to establish the strength of the relationship by calculating what is called the correlation coefficient. There are several types of correlation coefficients but the best known and most used is the Pearson's (after Karl Pearson) product moment correlation coefficient, r, for a sample and ρ (rho) for the population. The coefficient measures the strength and direction of a linear relationship between two variables. The coefficient lies between 0 and 1 for positive correlation and

between 0 and -1 for negative correlation, that is $0 \leq |r| \leq 1$. If there is perfect positive correlation r = +1.00; perfect negative correlation is indicated by r= -1.00; no relationship is shown by r= 0.00. For perfect negative correlation the scatter points form a straight line when plotted and the slope of the graph is negative. For a negative correlation the line of best fit is negatively sloped when plotted. On the other hand for perfect positive correlation the scatter points form a straight line when plotted and the slope of the graph is positive. For a positive correlation the line of best fit is positively sloped when plotted.





Correlation coefficient (ρ) Corr(X,Y)=

$$\frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E[(X-\mu_x)^2(Y-\mu_y)^2]}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) = \rho$$

Sample correlation coefficient r =Corr(X,Y)=

$$\frac{Cov(X,Y)}{S_x S_y} = \frac{S_{x,y}}{S_x S_y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S_x}\right)\left(\frac{Y_i-\overline{Y}}{S_y}\right) = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = r$$

or

$$r = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{(n-1)(s_x)(s_y)}$$

or

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

or

$$r = \frac{\sum [(Y_i-\overline{Y})(X_i-\overline{X})]}{\sqrt{[\sum (Y_i-\overline{Y})^2][\sum (X_i-\overline{X})^2]}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

or

$$r = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\sum Y^2 - n\overline{Y}^2}\sqrt{\sum X^2 - n\overline{X}^2}}$$ (Care should be taken here not to use the approximations for the two means).

- **Coefficient of Determination**

This measures the proportion of the total variation in the dependent variable that is explained by the variation in the independent or explanatory variables in the regression.[20] It is represented by r2 and varies between 0 and 1. An r2 value of 1.00 indicates that the regression equation "explains" 100 percent of the variation in the dependent variable about its mean. An r2 value in the 0.5 –1.00 range are usually interpreted to mean that the regression equation does a good job of explaining the Y variation.

$$r^2 = \frac{(total\ variance\ in\ the\ dependent\ variable) - (variance\ "unexplained"\ by\ regression\ equation)}{Total\ variance\ in\ the\ dependent\ variable}$$

We can look at r2 as Explained Variation in all items/ Total variation in all items.

$$r^2 = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{Explained\ Variation\ of\ Y}{Total\ Variation\ of\ Y}$$

4.6.4 Regression Methods for Sales Forecasting

1. Linear Regression

Linear regression is the simplest and most widely used statistical technique for predictive modeling. It basically gives us an equation, where we have our features as independent variables, on which our target variable [sales in our case] is dependent upon[21].

So what does the equation look like? Linear regression equation looks like this:

$$Y = \theta_1 X_1 + \theta_2 X_2 + \ldots \theta_n X_n$$

Here, we have Y as our dependent variable (Sales), X's are the independent variables and all thetas are the coefficients. Coefficients are basically the weights assigned to the features, based on their importance. For example, if we believe that sales of an item would have higher dependency upon the type of location as compared to size of store, it means that sales in a tier 1 city would be more even if it is a smaller outlet than a tier 3 city in a bigger outlet. Therefore, coefficient of location type would be more than that of store size.So, firstly let us try to understand linear regression with only one feature, i.e., only one independent variable. Therefore our equation becomes,

$$Y = \theta_1 * X + \theta_0$$

This equation is called a simple linear regression equation, which represents a straight line, where 'Θ0' is the intercept; 'Θ1' is the slope of the line. Take a look at the plot below between sales and MRP.

❖ Important Points:

- There must be linear relationship between independent and dependent variables

- Multiple regression suffers from multicollinearity, autocorrelation, heteroskedasticity.
- Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.

2. Ridge Regression model

Ridge regression chooses parameter estimates, $\hat{\beta}^{ridge}$, to minimize the residual sum of squares subject to a penalty on the size of the coefficients. After standardizing all potential terms in the model the ridge coefficients minimize [22]

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_o - \sum_{j=1}^{k} u_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right\}$$

Here l > 0 is a complexity parameter that controls the amount of shrinkage, the larger l the greater the amount of shrinkage. The intercept is not included in the shrinkage and will be estimated as the mean of the response. An equivalent way to write the ridge regression criterion is

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_o - \sum_{j=1}^{k} u_{ij}\beta_j)^2 \right\} \; subject\ to \sum_{j=1}^{k} \beta_j^2 \leq t$$

This clearly shows how of the size of the parameter estimates are constrained. Also this formulation of the problem also leads to a nice geometric interpretation of how the penalized least squares estimation works (see figure next page).

❖ Important Points:

- The assumptions of this regression is same as least squared regression except normality is not to be assumed
- It shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection feature
- This is a regularization method and uses l2 regularization.

3. Lasso Regression

The lasso is another shrinkage method like ridge, but uses an L1-norm based penalty.

The parameter estimates are chosen according to the following [22]

$$\hat{\beta}^{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_o - \sum_{j=1}^{k} u_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{k}\left|\beta_j\right|\right\}$$

$$= \min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_o - \sum_{j=1}^{k} u_{ij}\beta_j)^2\right\} \; subject\ to \sum_{j=1}^{k}\left|\beta_j\right| \le t$$

Here t > 0 is the complexity parameter that controls the amount of shrinkage, the smaller t the greater the amount of shrinkage. As with ridge regression, the intercept is not included in the shrinkage and will be estimated as the mean of the response. If t is chosen larger than $t_o = \sum_{j=1}^{k}\left|\hat{\beta}_j^{ls}\right|$ then there will be no shrinkage and the lasso estimates will be the same as the OLS estimates. If $t = t_o/2$ then the OLS estimates will be shrunk by about 50%, however this is not to say that $\widehat{\beta_j}^{lasso} = \hat{\beta}_j^{ls}/2$. The shrinkage can result in some parameters being zeroed (which ridge will not ever do), essentially dropping the associated predictor from the model as the figure below shows. In the diagram below, the lasso estimate for $\hat{\beta}_1^{lasso} = 0$.

Important Points:

- The assumptions of this regression is same as least squared regression except normality is not to be assumed
- It shrinks coefficients to zero (exactly zero), which certainly helps in feature selection
- This is a regularization method and uses l1 regularization

- If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero

4. Decision Tree Regression

The method starts by searching for every distinct value of all its predictors,[23] and splitting the value of a predictor that minimizes the following statistic (other regression tree models have different optimization criteria):

$$SSE = \sum_{i \in S_1} (y_i - \overline{y_1}) + \sum_{i \in S_2} (y_i - \overline{y_2})$$

Where $y1^{-}$ and $y2^{-}$ are the average values of the dependent variable in groups S1 and S2. For groups S1 and S2, the method will recursively split the predictor values within groups. In practice, the method stops when the sample size of the split group falls below certain threshold e.g. 50. To prevent over-fitting, the constructed tree can be pruned by penalizing the SSE (Sum of Squared Error) with tree size: $\boldsymbol{SSEcp = SSE + cp \times St}$ Where St is the size of the tree (number of terminal nodes), and cp is complexity parameter. Smaller cp will lead to larger trees, and vice versa. Of course, this parameter can also be tuned by cross-validation. Unlike linear regression models that calculate the coefficients of predictors, tree regression models calculate the relative importance of predictors. The relative importance of predictors can be computed by summing up the overall reduction of optimization criteria like SSE.

5. ADABoost Regression

An AdaBoost regressor[23] is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.and there by modifies the diversity of the training data. A small threshold will classify the majority of predicted values as incorrect and result in a near uniform sampling distribution, which will resemble bagging. A threshold that is set too large will result in fitting to extreme outliers. An adaptive work-around for the problem has been proposed by recursive adjustment of φ. However, the equations presented i

enable the value of φ to become negative, which is physically meaningless and force the algorithm to fail. The value of φ at iteration t+1 is

$\varphi_{t+1} = \varphi_t(1 - \lambda)$, while $e_t < e_{t-1}$

$\varphi_{t+1} = \varphi_t(1 + \lambda)$, while $e_t > e_{t-1}$

Where et is the root mean squared error (RMSE) at iteration t and by definition $e_t \geq 0$. The value λ is

$$\lambda = r \left| \frac{e_t - e_{t-1}}{e_t} \right|$$

where r is a tuning constant. The term 1-λ will be negative, leading to a negative φ, if $e_{t-1} > 2e_t$ (for r=1). This problem can be managed by selecting the value of r sufficiently small to let $\lambda < 1$. The reduction of the magnitude of r also limits the algorithm's ability to change and adapt the value of φ. The method shifts the tuning parameter from φ to r. The adaptive method will however not fail as long as

$$\frac{e_t}{r} > |e_t - e_{t-1}|$$

as long as $e_t \neq 0$ An additional limitation of Adaboost.RT is the definition of the misclassification function Ert(i) as the absolute relative error. This error function is a direct implementation of the Adaboost.M1 algorithm and classifies a sample prediction based on the percent error of the prediction ft(xi) compared to the true value yi . The absolute relative error is calculated at each iteration and is defined as

$$Er_t(i) = ARE_t(i) = \left| \frac{f_t(x_i) - y_i}{y_i} \right|$$

This definition of the error function contains a singularity when the true value yi=0. Values near zero receive an artificially high absolute relative error and are always classified as incorrect and their sampling weight increases. Values near zero will be selected more frequently for training and the accuracy of the values away from the zero-crossing will decrease. The performance of the entire ensemble is degraded when the output

6. Random Forest Regression

A random forest is a Meta estimator that fits a number of classifying[24] decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

4.7 Model Evaluation Method

How accurate do you think the model is? Do we have any evaluation metric, so that we can check this? Actually we have a quantity, known as R-Square.

1. R-Square: It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). Mathematically,[26] it can be written as:

$$R - Square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})\text{^}2}{\sum(Y_{actual} - Y_{mean})\text{^}2}$$

The value of R-square is always between 0 and 1, where 0 means that the model does not model explain any variability in the target variable (Y) and 1 meaning it explains full variability in the target variable.

2. Root Mean Square Error

The Root Mean Square Error (RMSE)[27] (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modeled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power. The RMSE of a model prediction with respect to the estimated variable Xmodel is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

where Xobs is observed values and Xmodel is modeled values at time/place i. The calculated RMSE values will have units, and RMSE for phosphorus concentrations can for this reason not be directly compared to RMSE values for chlorophyll a concentrations etc. However, the RMSE values can be used to distinguish model performance in a calibration period with that of a validation period as well as to compare the individual model performance to that of other predictive models.

CHAPTER: FIVE

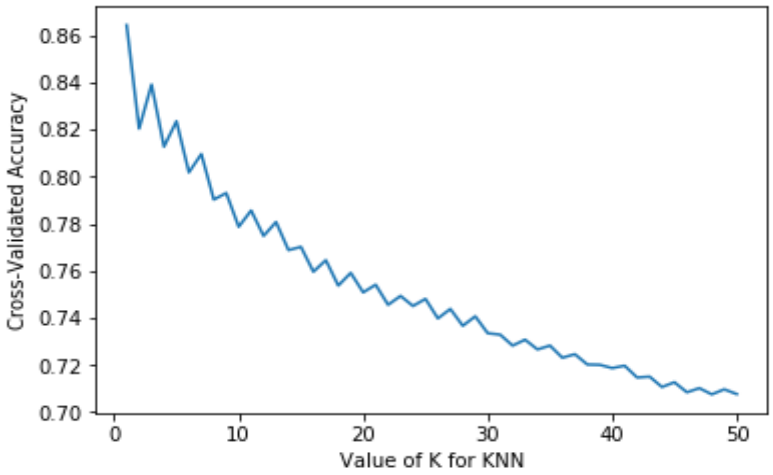
EXPERIMENTAL RESULTS AND DISCUSSIONS

5.1 Introduction

In this chapter we will add some output screenshots and discuss about this.

5.2 Experimental Result

```
Location for Max Accuaracy is:
1
Value of K with is: 1
```



We see that if K is increasing then Cross-Validated Accuracy decrease

Fig: Cross-Validation Accuracy vs. Value of K for KNN Output Diagram

Accuracy of our Dataset is given below:

```
Test Accuracy:  0.8519447929736512
```

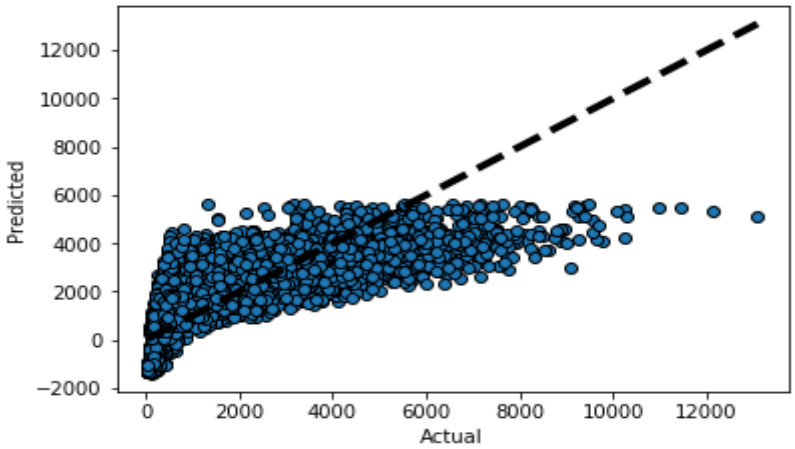


Fig: Linear Regression Model Output

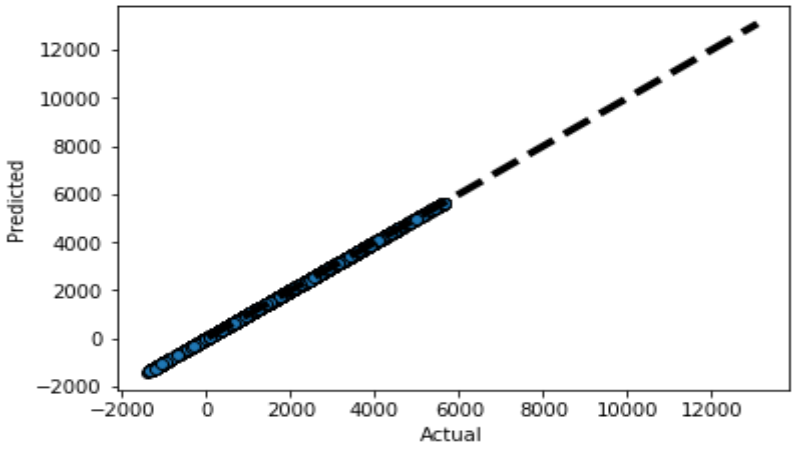


Fig: Ridge Regression Model Output

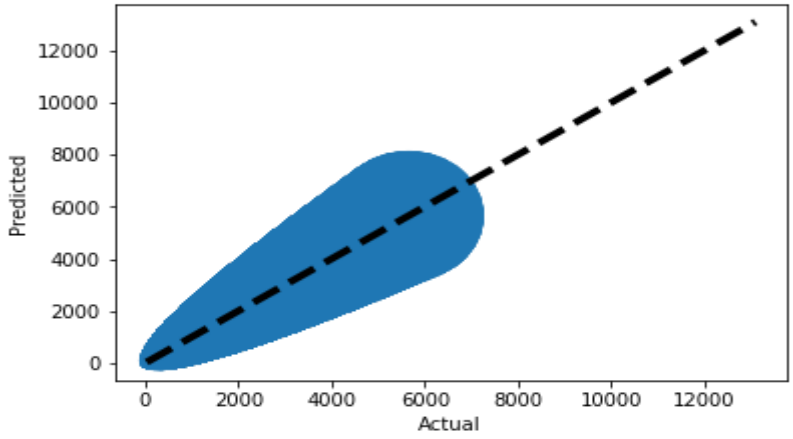


Fig: Lasso Regression Model Output

5.3 Result Evaluation

5.3.1 RMSE (Root mean squared error) and R squared (r2_score) Values

Here RMSE values indicate that difference between actual data to model's predicted values. Lower values of RMSE indicate better fit. RMSE can range from 0 to 1 .R-squared provides the relative measure of the percentage of the dependent variable variance that the model explains. Higher R-squared values indicate that the data points are closer to the fitted values. R-squared can range from 0 to 100%.

1. Linear Regression Model RMSE and r2_score values

```
Out[277]: 48.70715949972778
```

```
Out[278]: 0.4180537250203177
```

2. Lasso regression RMSE and r2_score values

```
48.656871372516605
0.41865456047462923
```

3. AdaBoost Regression RMSE and r2_score values

```
48.700028211358386
0.418138928648388
```

4. Ridge Regression RMSE and r2_score values

```
48.69851612158346
0.41815699488384483
```

5. Decision Tree Regression RMSE and r2_score values

```
48.865410192892206
0.4161629680485319
```

6. Random Forest Regression RMSE and r2_score values

```
48.44411027965156
0.4211966000171756
```

Here we see that every model have quite similar RMSE and R squared values but among them decision tree has the highest percentage of R squared and lowest RMSE values which indicates that data points are closer to the fitted values.

Out[301]:

	Item_Identifier	Outlet_Identifier	Item_Outlet_Sales
0	FDW58	OUT049	1778.765377
1	FDW14	OUT017	1396.876176
2	NCN55	OUT010	599.071456
3	FDQ58	OUT017	2591.601758
4	FDY38	OUT027	6212.526922

Fig: Combining all models final output

From this output snap we see that how many times any specific item can be sold in a particular outlet on upcoming month.

CHAPTER: SIX

CONCLUSION

6.1 Result and Discussion

This research took us through the entire journey of solving a data science problem. We started with making some hypothesis about the data without looking at it. Then we moved on to data exploration where we found out some nuances in the data which required remediation. Next, we performed data cleaning and feature engineering, where we imputed missing values and solved other irregularities, made new features and also made the data model-friendly by one-hot-coding. Finally we made regression, decision tree, and random forest model and got a glimpse of how to tune them for better results.

6.2 Limitations

- Part hard fact, part guesswork
- Forecast may be wrong
- Times may change

6.3 Future Works

There is also another popular Machine Learning model for sales forecasting called Time Series Analysis. In future I try to implement Time Series Analysis for better outcomes.

REFERENCE

[1] https://web.njit.edu/~turoff/pubs/delphibook/delphibook.pdf The Delphi Method Techniques and Applications Harold A. Linstone Portland State University Murray Turoff New Jersey Institute of Technology ©2002 Murray Turoff and Harold A. Linstone

[2] Armstrong, J. S. (1987). The forecasting audit. In: Makridakis, S. & Wheelwright, S. C. f (Eds.), The Handbook of Forecasting, John Wiley & Sons, New York, pp. 584-602.

[3] Diamantopoulos, A. & Winklhofer, H. (1998). A Conceptual Model of Export Sales Forecasting Practice and Performance: Development and Testing, In: Anderson, P. (Ed.), Proceedings of the 27th European Marketing Acadeniv Conference (May. Stockholm, Sweden), pp. 57-83.

[4] Jones, V. S., Bretschneider, S., & Gorr, W. (1997). Organizational pressures on forecast evaluation: managerial, political, and procedural influences. Journal of Forecasting, 16, 241-254.

[5] Mahmoud, E., DeRoeck, R., Brown, R. G., & Rice, G. (1992). Bridging the gap between theory and practice in forecasting. International Journal of Forecasting, 8, 251-267.

[6] Winklhofer, H., Diamantopoulos, A., & Witt, S. F. (1996). Forecasting practice: a review of the empirical literature and an agenda for future research. International Journal of Forecasting, 12, 193-221.

[7] Researching Sales Forecasting Practice Commentaries and authors' response on "Conducting a Sales Forecasting Audit" by M.A. Moon, J.T. Mentzer & C.D. Smith International Journal of Forecasting 19 (2003) 27-42

[8] Forecasting Methods for Marketing: Review of Empirical Research Published in International Journal of Forecasting, 3 (1987), 355-376. J. Scott Armstrong The Wharton School, University of Pennsylvania Roderick J. Brodie University of Canterbury, New Zealand Shelby H. McIntyre Santa Clara University, CA

[9] Best, Roger J., 1974, an experiment in Delphi estimation in marketing decision making, Journal of Marketing Research 11, 448-452.

[10] Brodie, Roderick J. and C. A. de Kluyver, 1984, Attraction versus linear and multiplicative market share models, Journal of Marketing Research 21, 196-201. Brodie, Roderick J. and C. A. de Kluyver, 1987, A comparison of the short-term forecasting accuracy

of econometric and naive extrapolation models of market share, International Journal of Forecasting 3, 423-437.

[11] John G. Wacker, Rhonda R. Lummus, (2002) "Sales forecasting for strategic resource planning", International Journal of Operations & Production Management, Vol. 22 Issue: 9, pp.1014-1031, https://doi.org/10.1108/01443570210440519

[12] Developments in Business Simulation and Experiential Learning, Volume 31, 2004 online sales forecasting with the multiple regression analysis data matrices package

[13] Enrick, N. L. (1969) Market and Sales Forecasting: A Quantitative Approach, San Francisco, CA: Chandler.

[14] Faria, A. J., Nulsen, Jr., R. O., & Roussos, D. S. (1994), COMPETE: A Dynamic Marketing Simulation, 4th ed. Burr Ridge, IL: Irwin.

[15] Guerard, J.B., Jr. and E. Schwartz, 2007. Quantitative Corporate Finance. New York: Springer. Gunst, R.F. and R.L. Mason, 1980. Regression Analysis and its Application, New York: Marcel Dekker, Inc

[16] Cochrane, D. and G.H. Orcutt. 1949. Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. Journal of the American Statistical Association. 44:32-61.

[17] Cook, R.D. 1977. “Detection of Influential Observations in Linear Regression.” Technometrics 19, 15–18.

[18] Cochrane, D. and G.H. Orcutt. 1949. Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. Journal of the American Statistical Association. 44:32-61

[19] Gunst, R.F. and R.L. Mason, 1980. Regression Analysis and its Application, New York: Marcel Dekker, Inc.

[20]https://www.researchgate.net/publication/267093056_Applied_Regression_Analysis_and_Other_Multi-Variable_Methods

[21] https://www.kaggle.com

[22] https://towardsdatascience.com/data-science-case-study-optimizing-product-placement-in-retail-part-1-2e8b27e16e8d

[23] https://en.wikipedia.org/wiki/Machine_learning

[24] https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

[25],https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf

[26] http://statisticsbyjim.com/regression/standard-error-regression-vs-r-squared/

[27] http://statisticsbyjim.com/regression/standard-error-regression-vs-r-squared/

[28]https://www.analyticsinsight.net/intuition-behind-bias-variance-tradeoff-lasso-and- ridge-regression/

SCREENSHOTS

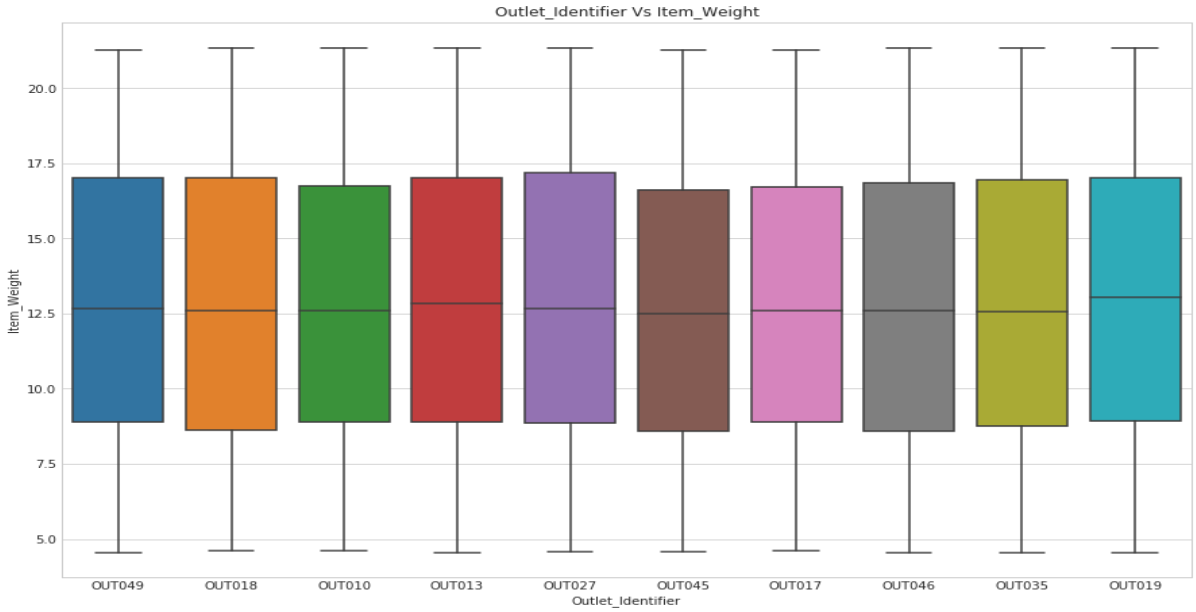


Fig: Outlet_Identifier Vs. Item_Weight

Looking at those plots we will notice that all the Outlet Identifier (Unique store ID) have identical median, boxes and whiskers.

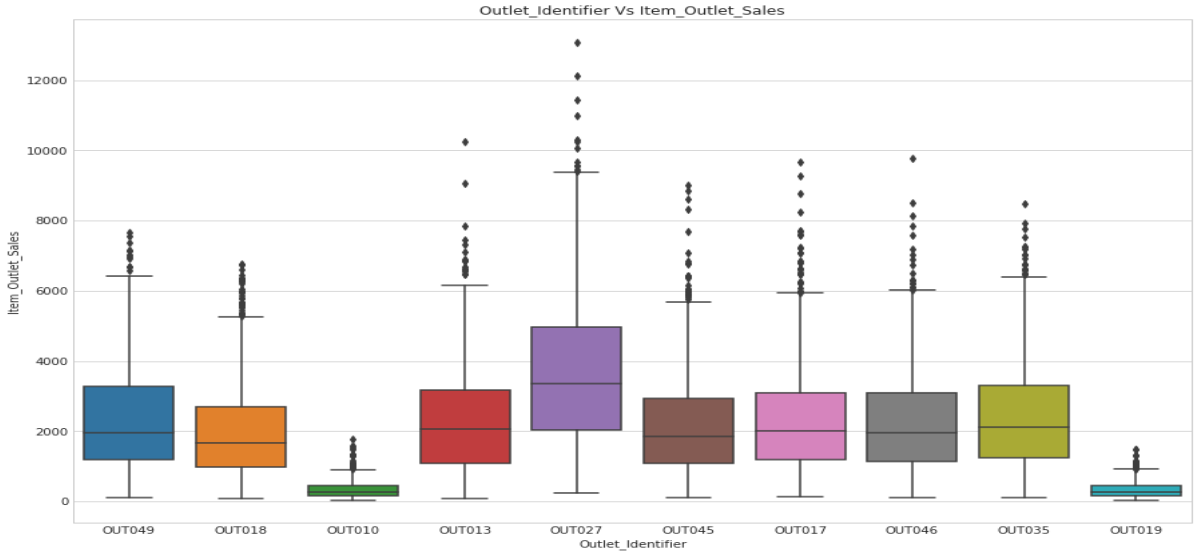


Fig: Outlet_Identifier Vs. Item_Outlet_Sales

From the above boxplot, we see that the two "Grocery stores" OUT010 and OUT019 have reported far fewer sales than the supermarkets.

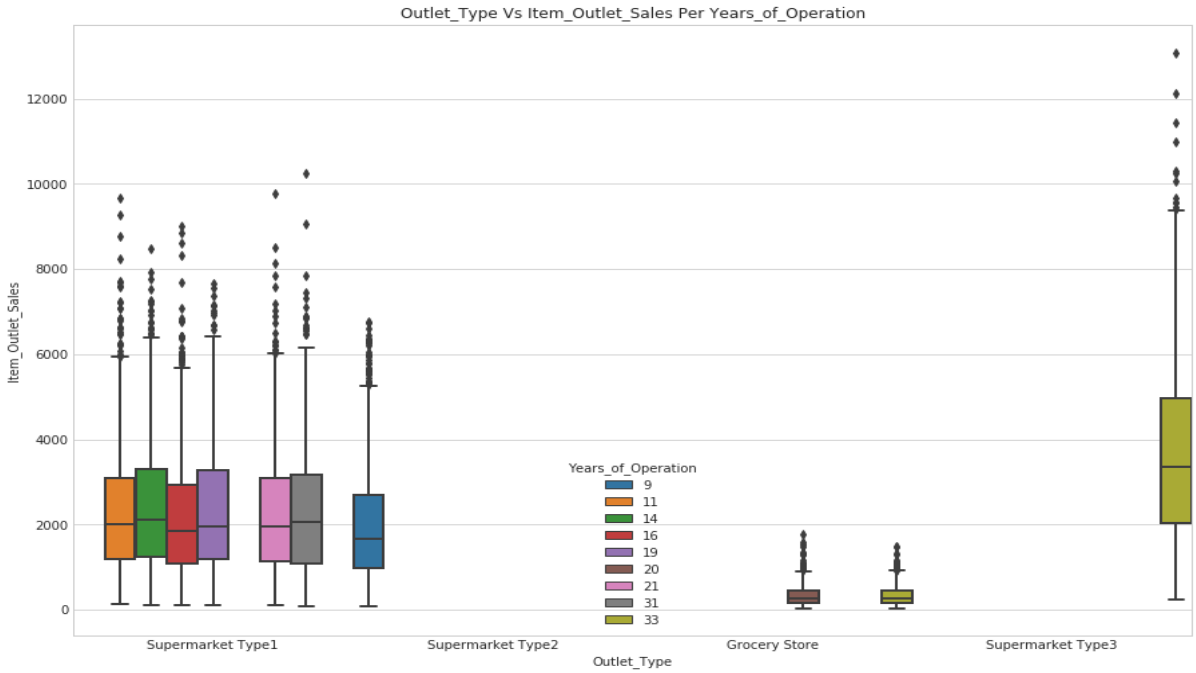


Fig: Outlet_Type Vs Item_Outlet_Sales per Years_of_Operation

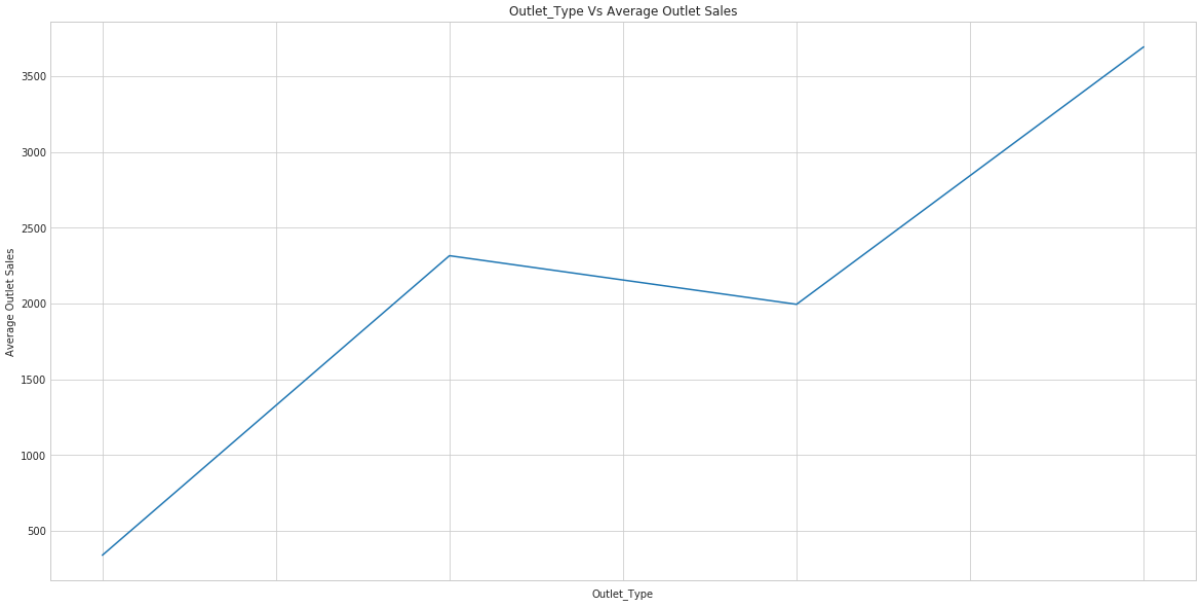


Fig: Outlet_Type Vs. Average Outlet Sales

Grocery stores have the least average sales while Supermarket Type 3 have the most average sales

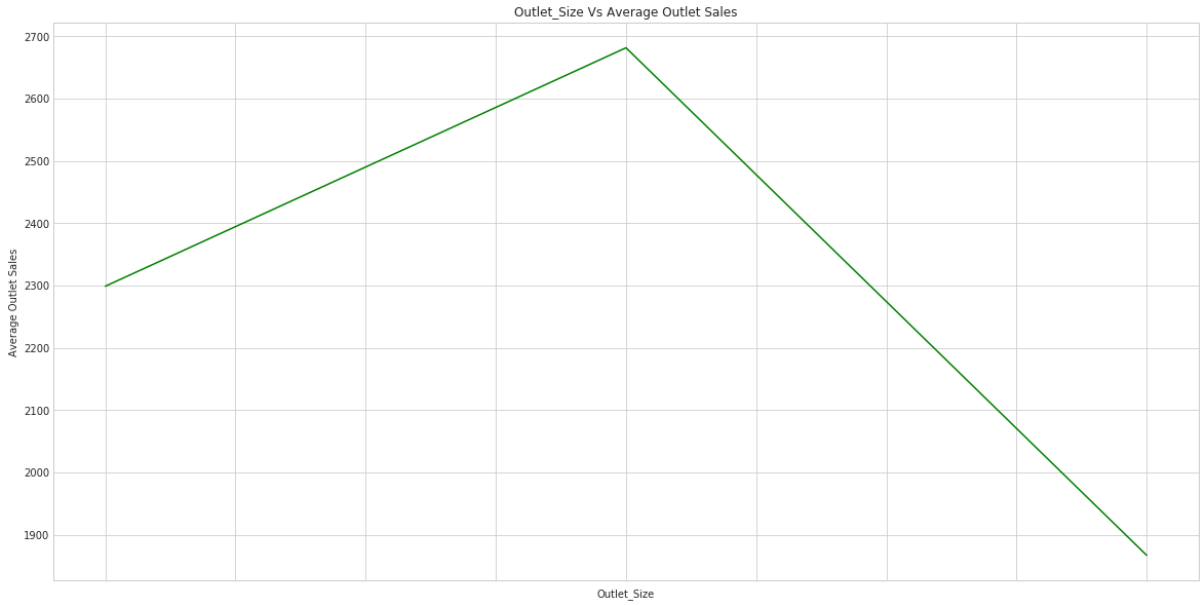


Fig: Outlet_Size vs. Average Outlet Sales

In this diagram show we see that average outlet sales decrease or increase on the basis of outlet size

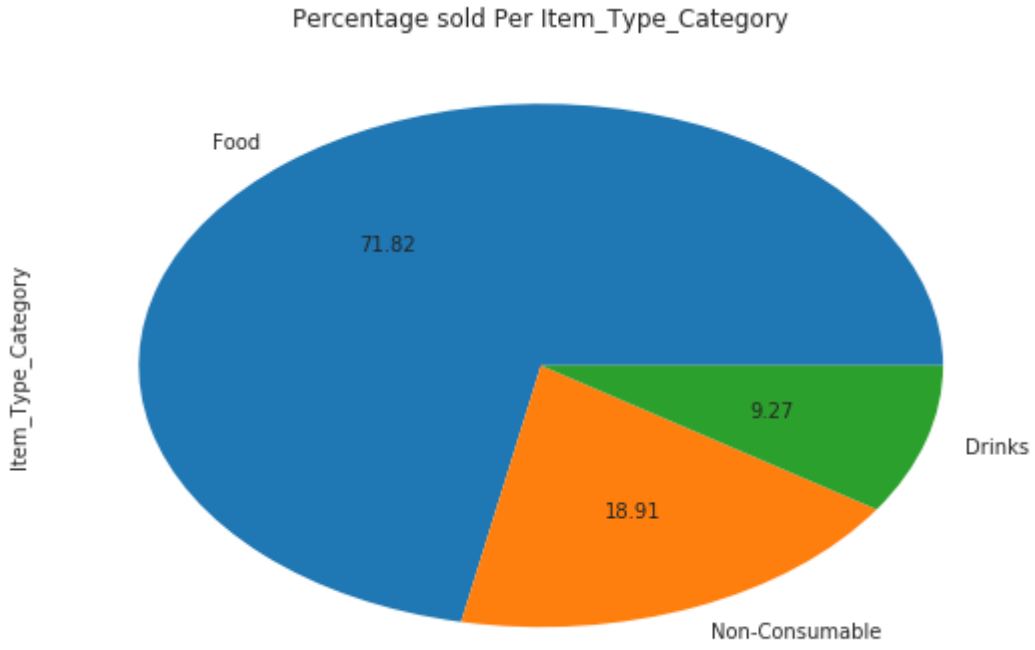


Fig: Percentage sold Per Item_Type_Category

The majority (71.82%) of items sold is classified as Food

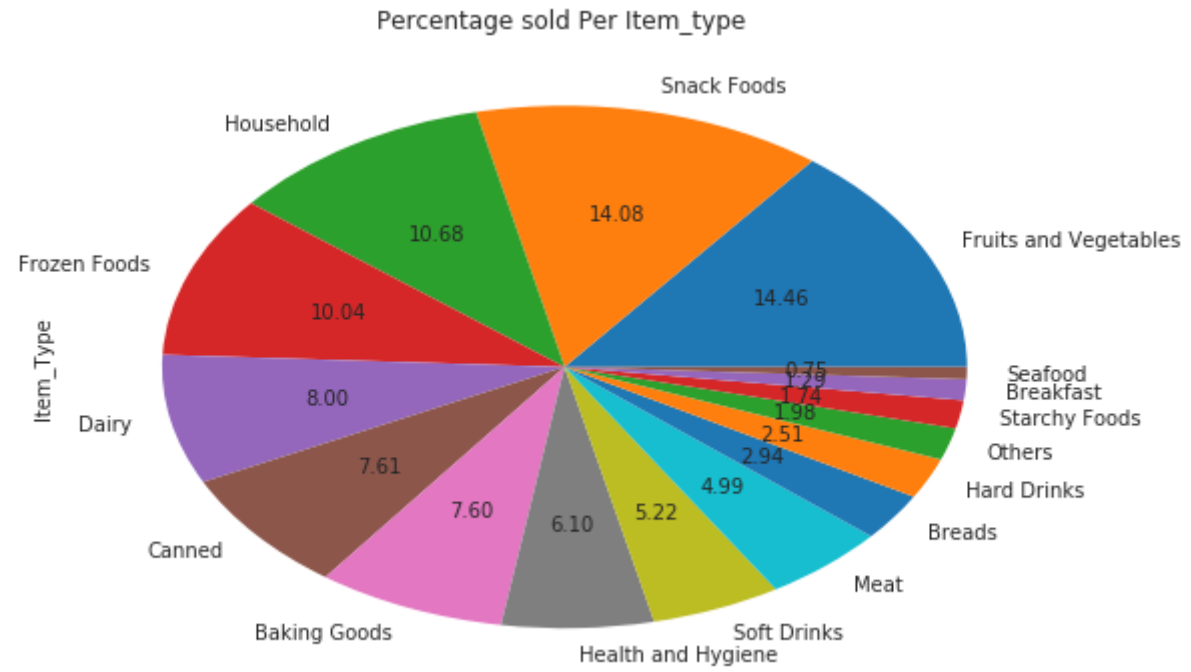


Fig: Percentage sold Per Item_type

Item type with the highest sales (14.46% of all sales) is Fruits and Vegetables with Snack and Foods (14.08) a close second. The item with the least sales (0.75%) is seafood.

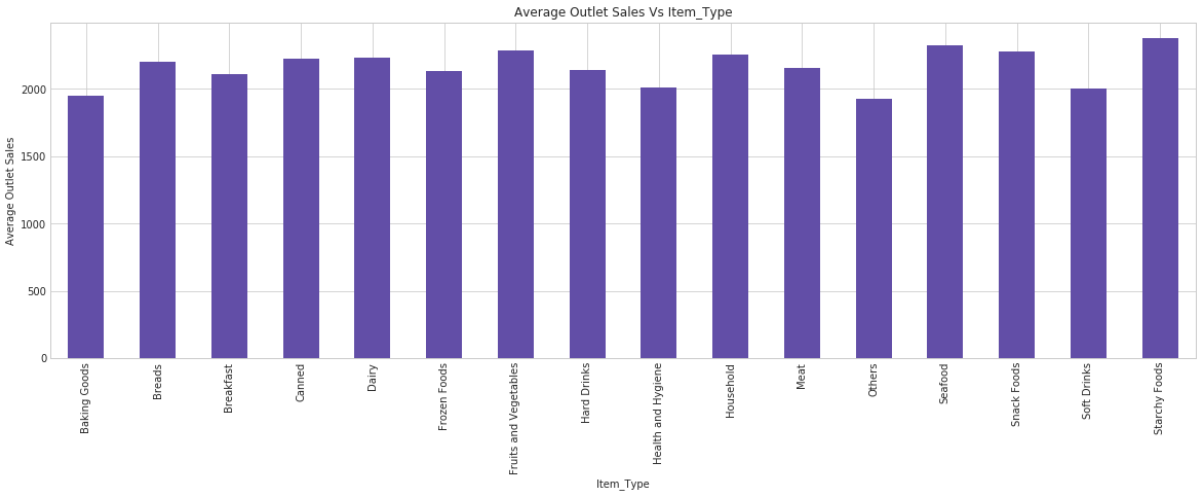


Fig: Average Outlet Sales Vs Item_Type

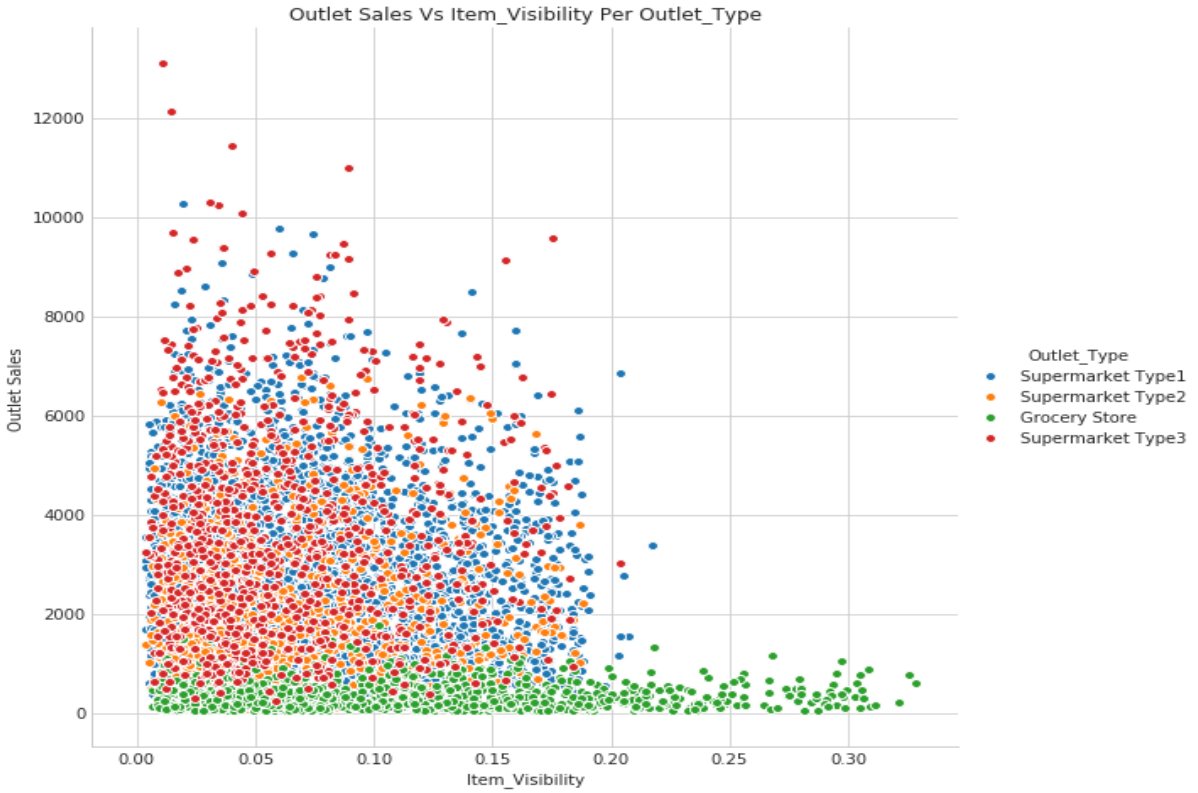


Fig: Outlet Sales Vs Item_Visibility Per Outlet_Type

Again, we can clearly see the poor sales performance of items from Grocery stores

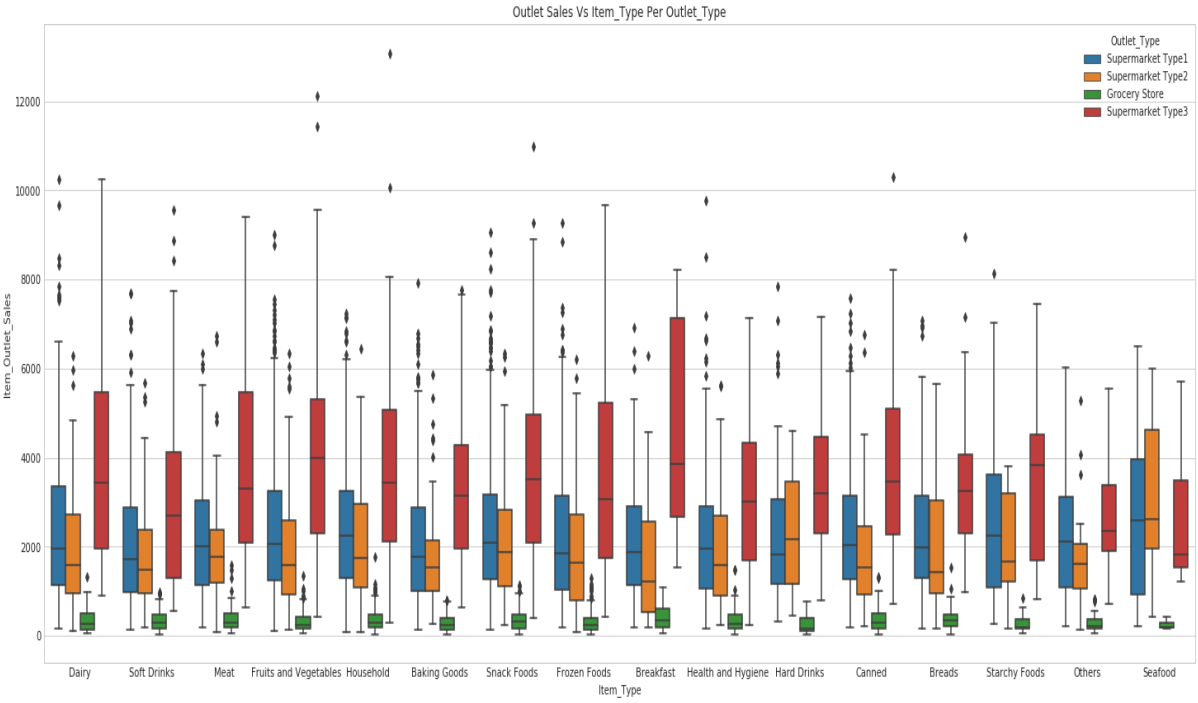


Fig: Outlet Sales Vs Item_Type per Outlet_Type

Supermarket Type3 seems to consistently outperformed other store outlets in all item types

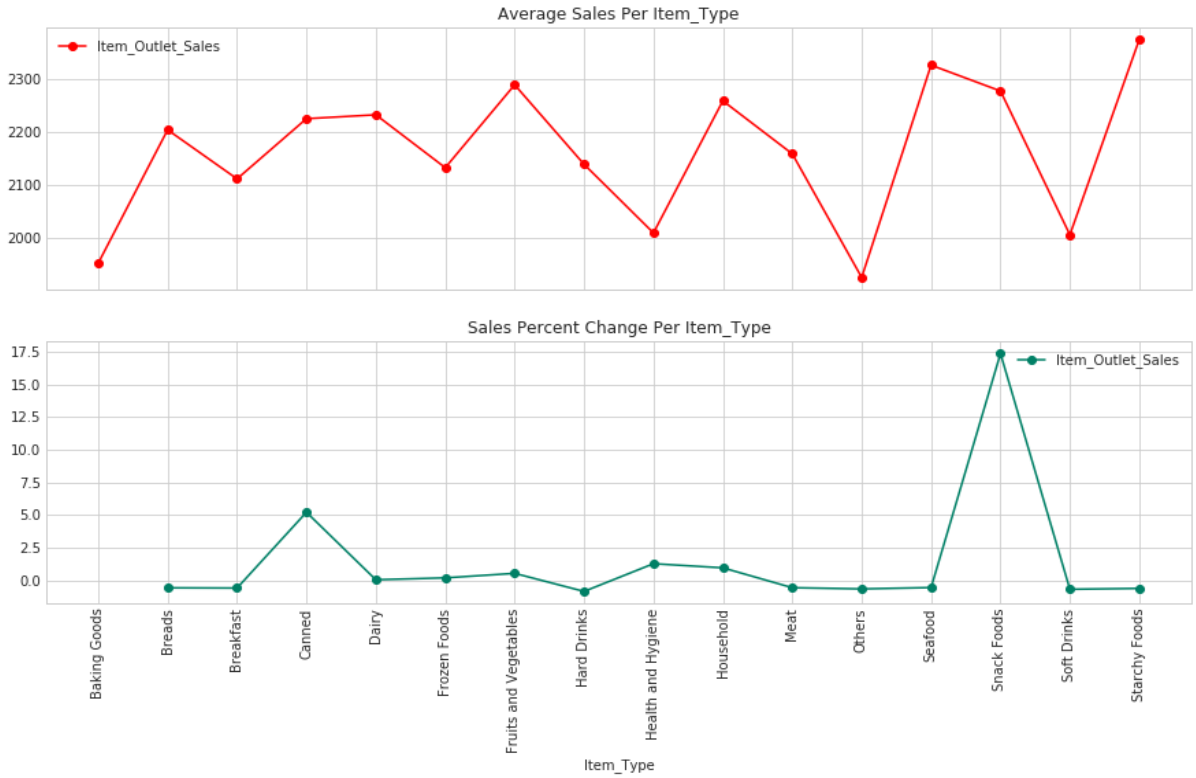


Fig: Average Sales per Item_Type Sales Percent Change per Item_Type