APPENDIX 1

SALES FORECASTING IN BIG MART

A Project Report of Capstone Project - 2

Submitted by

RAHUL DWIVEDI

(1613112035/16SCSE112042)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION IN DATA ANALYTICS

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

Dr. SATYAJEE SRIVASTAVA, Ph. D.,

Associate Professor

APRIL/MAY - 2020

DECLARATION

Project Title: **Sales Forecasting In Big Mart**

Degree for which the project work is submitted: **Bachelor of Technology in Computer Science and Engineering**

I declare that the presented project represents largely my own ideas and work in my own words. Where others ideas or words have been included, I have adequately cited and listed in the reference materials. The report has been prepared without resorting to plagiarism. I have adhered to all principles of academic honesty and integrity. No falsified or fabricated data have been presented in the report. I understand that any violation of the above will cause for disciplinary action by the Institute, including revoking the conferred degree, if conferred, and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken.

Rahul Dwivedi
(1613112035)

Date:

APPENDIX 2

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "SALES FORECASTING IN BIG MART" is the bonafide work of "RAHUL DWIVEDI(1613112035)" who carried out the project work under my supervision.

SIGNATURE OF HEAD
Dr. MUNISH SHABARWAL,
Ph. D. (Management),Ph. D.(CS)
Professor & Dean,
School of Computer Science & Engineering

SIGNATURE OF SUPERVISOR
Dr. SATYAJEE SRIVASTAVA
Ph. D.(CS)
Associate Professor,
School of Computer Science & Engineering

ABSTRACT

Sales forecasting is one of the most important needs of the retail industry .The retailers face a lot of new challenges with the increasing competition with e-commerce sites as well as with other ever-changing dynamics of the market. So to analyze the data and predict sales of certain products and their overall impact on the sales of certain stores we use several machine learning algorithms. The data we use here is 2013 Sales data of Big Mart .We use various techniques like linear regression, ridge regression which are the basic machine learning algorithm and test their accuracy with those of the more newer methods like random forests and with that of methods which can now be run with the parallel systems like Xgboost which is the more rigorous form of GBM. The objective of the paper is to check the overall efficiency of the systems and accuracy of the results.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my guide, Dr. Satyajee Srivastava for his valuable guidance, consistent encouragement and providingus with an excellent atmosphere for doing research. He has extended cordial support to us for completing our project.

Rahul Dwivedi

APPENDIX 3

TABLE OF CONTENTS

APPENDIX 1

DECLARATION

APPENDIX 2

ABSTRACT..... iii

ACKNOWLEDGEMENT..... iv

LIST OF FIGURES……..5

LIST OF ABBREVIATIONS……6

CHAPTER - 1.....7
INTRODUCTION
(i) Sales……7
(ii) Forecasting……...7
(iii) Motivations and scope……..……..8

CHAPTER - 2
FORECASTING METHODS…..……9

CHAPTER – 3……16
1) **CAUSAL APPROACH TO SALES FORECASTING……..16**
2) Factors affecting sales……17

CHAPTER-4……21
NEW PRODUCT FORECASTING……..……..…..21

CHAPTER - 5……..30
PROPOSED MODEL…….30

CHAPTER - 6……..37
IMPLEMENTATION AND ARCHITECTURE DESIGN……37

CHAPTER – 7…….……51
RESULTS AND DISCUSSION……..51

CHAPTER - 8……55
CONCLUSION AND FUTUTRE WORK……..……..55

CHAPTER – 9……..……..57

REFERENCES……57

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1	Different Forecasting Techniques	
2	Details of forecasting techniques	
3	Causal approach to sales forecasting	
4	Different errors in surveying	
5	Multiple linear regressions	
6	Node information in decision trees	
7	Ridge regression graph	
8	Random forest explanation	
9	Best model parameters	
10	XGboost Explanation	

11 Flowchart for sales forecasting via different techniques

12 Target Variable

13 Uni variate Analysis 1

14 Uni variate analysis 2

15 Uni variate analysis 3

16 Uni variate analysis 4

17 Variable Importance

18 Different algorithm performance

LIST OF ABBREVIATIONS

B2C	Business to consumer
EDA	Exploratory Data Analysis
ML	Machine Learning
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
XGboost	Extreme Gradient Boosting

CHAPTER-1
INTRODUCTION

1. Sales:

Sales is the record of the amount of items sold of a particular product or service. Companies use various models to predict the amount of sales of the product but smaller business does'nt have enough capital in order to predict the data .These sales help he business owners determine the amount it requires to gain a certain amount of profit while leaving some for the retailers and the middle men. It helps in accounting the amount of product sold and what is the profit percentage.

2. Forecasting:

Forecasting is the process of making predictions about the future based on the past and present data. Here we predict the sales of the big mart 2013 data. Sound predictions and forecasting are must these days as
the companies have to adjust to the production of the product according to the various factors that affect the sales like the season ,sudden demands, price cut, competitive adaptability etc. To deal with different types of problems different forecasting techniques have been developed. The selection of the forecasting method depends on many factors like

- Accuracy of data
- Time period of the forecast
- Availability of the past data

These are some of the factors. The company should choose a technique to get the best use of available data. The forecast for a particular product requires to determine the stage of the lifecycle of the product i.e. its forecasting depends on the maturity of product. For

successful forecasting the company needs to answer the following questions

Purpose of forecast: Its purpose is to identify what business the company should enter in order to make profit. Simple sketch of the future is not good enough in order to predict the future .Different techniques vary in terms of cost, accuracy so in order to get best results in a certain range of cost a trade off is met. The other purpose of this is to identify the amount of product to make in order to avoid extra production which might not sell.

Dynamics and components of the system for which the forecast is made: this clarifies the relation of variables. A flowchart is made to review the position of the different positions of the different elements i.e. sales system, production system. This flowchart helps in determining the work dependency of variables.

Past data for future estimation: This is one of the most important step in future forecast many changes might not affect in the short term but they can affect in the long run. In order to make the accurate forecast of the sales we need to take old data as well as newer data in order to better forecast the future sales.

Sales forecasting: Sales forecasting involves predicting the amount people will purchase, given the product features and the conditions of the sale. Sales forecasts help investors make decisions about investments in new ventures. They are vital to the efficient operation of the firm and can aid managers on such decisions as the size of a plant to build, the amount of inventory to carry, the number of workers, to hire, the amount of advertising to place, the proper price to charge, and the salaries to pay salespeople. Profitability depends on

- Having a relatively accurate forecast of sales and costs;
- Assessing the confidence one can place in the forecast; and
- Properly using the forecast in the plan.

(i) Motivations and scope:

As with any machine learning task, data science specialists first need data to work with. Depending on the goal, researchers define what data they must collect. Next, selected data is prepared, preprocessed, and transformed in a form suitable for building machine learning models. Finding the right methods to training machines, fine-tuning the models, and selecting the best performers is another significant part of the work. Once a model that makes predictions with the highest accuracy is chosen, it can be put into production.

The overall scope of work data scientists carry out to build ML-powered systems capable to forecast customer attrition may look like the following:

- Understanding a problem and final goal
- Data collection
- Data preparation and preprocessing
- Modeling and testing

CHAPTER – 2

FORECASTING METHODS

Forecasting Methods: Forecasting involves methods that derive primarily from judgmental sources versus those from statistical sources. These methods and their relationships are shown in the flow chart in Figure 1. Judgment and statistical procedures are often used together, and since 1985, much research has examined the integration of statistical and judgmental forecasts (Armstrong and Collopy 1998b). Going down the figure, there is an increasing amount of integration between judgmental and statistical procedures. A brief description of the methods is provided here. Makridakis, Wheelwright and Hyndman (1998) provide details on how to apply many of these methods.

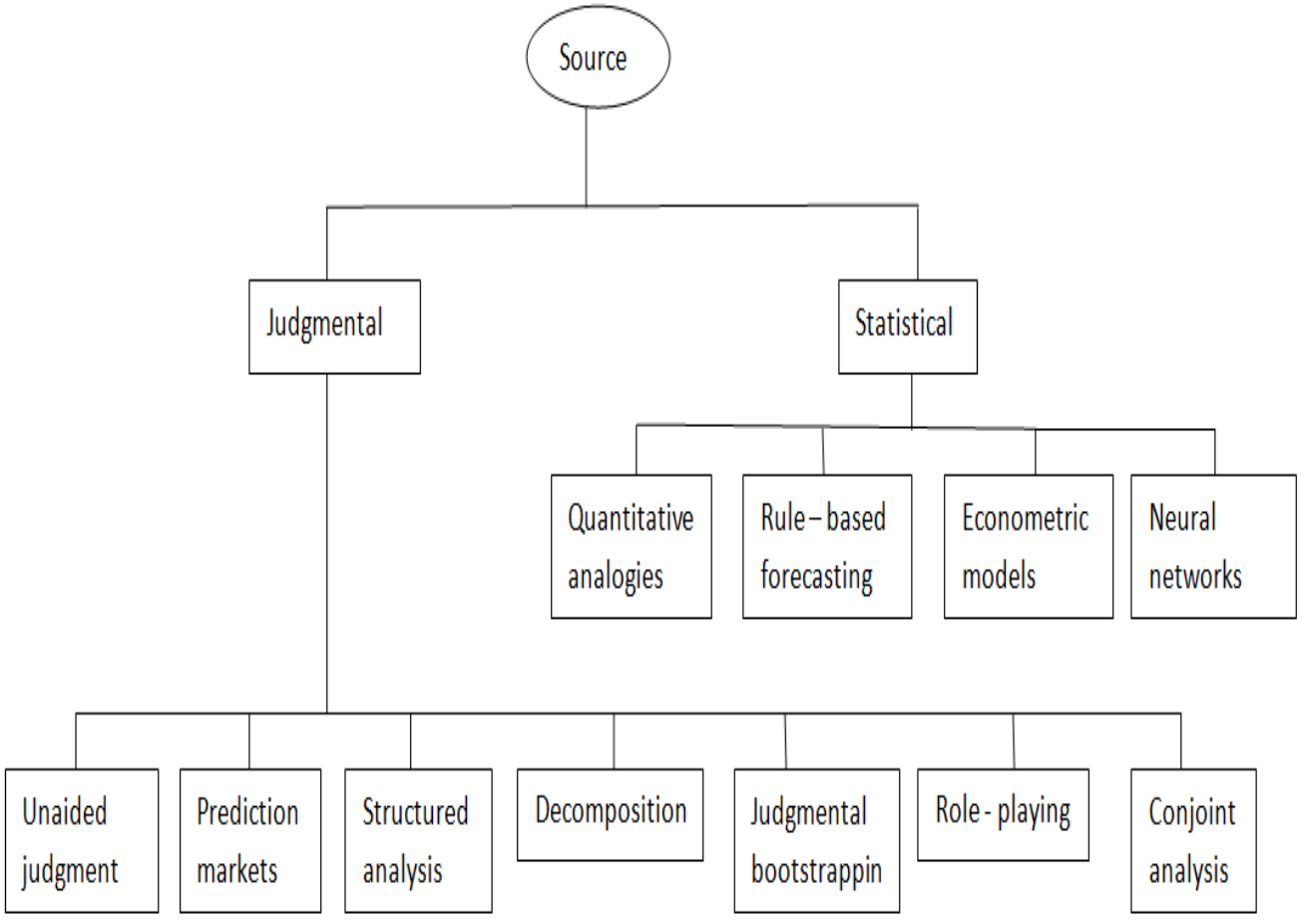


Figure 1: Different Forecasting Techniques

Intentions studies ask people to predict how they would behave in various situations. This method is widely used and it is especially important where one does not have sales data, such as for new product forecasts.

A person's role may be a dominant factor in some situations, such as in predicting how someone would behave in a job related situation. Role-playing is useful for making forecasts of the behavior of individuals who are interacting with others, and especially in situations involving conflict.

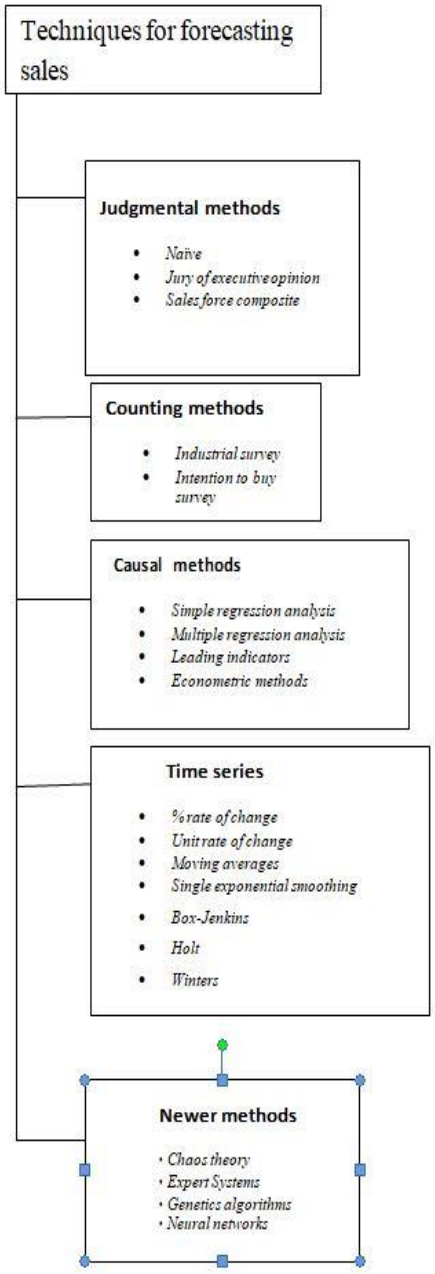


Figure 2 – Details of forecasting techniques

Another way to make forecasts is to ask experts to predict how others will behave in given situations. The accuracy of expert forecasts can be improved through the use of structured methods, such as the Delphi procedure. Delphi is an iterative survey procedure in which experts provide forecasts for a problem, receive anonymous feedback on the forecasts made by other experts, and then make another forecast. For a summary of the evidence on the accuracy of Delphi versus unstructured judgment, see Rowe and Wright (1999). One principle is that experts" forecasts should generally be independent of one another. Focus groups always violate this principle. As a result, they should not be used in forecasting.

Intentions can be explained by relating the "predictions" to various factors that describe the situation. By asking consumers to state their intentions to purchase for a variety of different situations, it is possible to infer how the factors relate to intended sales. This is often done by regressing their intentions against the factors, a procedure known as "conjoint analysis."

As with conjoint analysis, one can develop a model of the expert. This approach, judgmental bootstrapping, converts subjective judgments into objective procedures. Experts are asked to make a series of predictions. For example, they could make forecasts for the next year's sales in geographical regions. This process is then converted to a set of rules by regressing the forecasts against the information used by the forecaster. Once developed, judgmental bootstrapping models offer a low-cost procedure for making forecasts. They almost always provide an improvement in accuracy in comparison to judgmental forecasts, although these improvements are typically modest (Armstrong 1999).

Extrapolation methods use only historical data on the series of interest. The most popular and cost effective of these methods are based on exponential smoothing, which implements the useful principle that the more recent data are weighted more heavily. Another principle for extrapolation is to use long time-series when developing a forecasting model. Yet, Focus Forecasting, one of the most widely-used time-series

methods in business firms, does not do this. As a result, its forecasts are inaccurate (Gardner and Anderson 1997).

Still another principle for extrapolation is to use reliable data. The existence of retail scanner data means that reliable data can be obtained for existing products. Scanner data are detailed, accurate, timely and inexpensive. As a result, the accuracy of the forecasts should improve, especially because of the reduction in the error of assessing the current status. Not knowing where you are starting from has often been a major source of error in predicting where you will wind up. Scanner data are also expected to provide early identification of trends.

Empirical studies have led to the conclusion that relatively simple extrapolation methods perform as well as more complex methods. For example, the Box-Jerkins procedure, one of the more complex approaches, has produced no measurable gains in forecast accuracy relative to simpler procedures (Makridakis et al. 1984; Armstrong 1985). Although distressing to statisticians, this finding should be welcome to managers.

Quantitative extrapolation methods make no use of managements' knowledge of the series. They assume that the causal forces that have affected a historical series will continue over the forecast horizon. The latter assumption is sometimes false. When the causal forces are contrary to the trend in the historical series, forecast errors tend to be large (Armstrong and Collopy 1993). While such problems may occur only in a small minority of cases in sales forecasting, their effects can be disastrous. One useful guideline is that trends should be extrapolated only when they coincide with managements" prior expectations.

Judgmental extrapolations are preferable to quantitative extrapolations when there have been large recent changes in the sales level and where there is relevant knowledge about the item to be forecast (Armstrong and Collopy 1998b). Quantitative extrapolations have an advantage over judgmental methods when the large (Armstrong 1985, 393-401). More important than these small gains in accuracy, however, is that the quantitative methods are often less expensive. When one has thousands of forecasts to

make every month, the use of judgment is seldom cost effective.

Experts can identify analogous situations. Extrapolation of results from these situations can be used to predict for the situation that is of interest. For example, to assess the loss in sales when the patent protection for a drug is removed, one might examine the results for previous drugs. Incidentally, the first year loss is substantial.

Rule-based forecasting integrates judgmental knowledge about the domain. Rule-based forecasting is a type of expert system that is limited to statistical time series. Its primary advantage is that it incorporates the manager's knowledge in an inexpensive way.

Expert systems use the rules of experts. In addition, they typically draw upon empirical studies of relationships that come from econometric models. Expert opinion, conjoint analysis, bootstrapping and econometric models can aid in the development of expert systems.

Despite an immense amount of research effort, there is little evidence that multivariate time-series provide any benefits to forecasting. As a result, these methods are not discussed here.

Econometric models use data to estimate the parameters of a model given various constraints. When possible which is nearly always in management problems, one can draw upon prior research to determine the direction, functional form, and magnitude of relationships. In addition, they can integrate expert opinion, such as that from a judgmental bootstrapping model. Estimates of relationships can then be updated by using time-series or cross-sectional data. Here again, reliable data are needed. Scanner data can provide data from low-cost field experiments where key features such as advertising or price are varied to assess how they affect sales. The outcomes of such experiments can contribute to the estimation of relationships. Econometric models can also use inputs from conjoint models. Econometric models allow for extensive integration of judgmental planning and decision making. They can incorporate the effects of marketing mix variables as well as variables representing key aspects of the

market and the environment. Econometric methods are appropriate when one needs to forecast what will happen using different assumptions about the environment or different strategies. Econometric methods are most useful when (1) strong causal relationships with sales are expected; (2) these causal relationships can be estimated; (3) large changes are expected to occur in the causal variables over the forecast horizon; and (4) these changes in the causal variables can be forecast or controlled, especially with respect to their direction. If any of these conditions does not hold (which is typical for short-range sales forecasts), then econometric methods should not be expected to improve accuracy.

CHAPTER – 3

CAUSAL APPROACHES TO SALES FORECASTING

Instead of extrapolating sales directly, one can forecast the factors that cause sales to vary. This begins with environmental factors such as population, gross national product (GNP) and the legal system. These affect the behavior of customers, competitors, suppliers, distributors and complementors (those organizations with whom you cooperate). Their actions lead to a market forecast. Their actions also provide inputs for the market share forecast. The product of the market forecast and the market share forecast yields the sales forecast.

The breakdown of the problems into the elements of Figure 2 may aid one's thinking about the sales forecasts. It is expected to improve accuracy (versus the extrapolation of sales) only if one has good information about each of the components and if there is a good understanding about how each relates to sales. If there is high uncertainty about any of the elements, it might be more accurate to extrapolate sales directly.

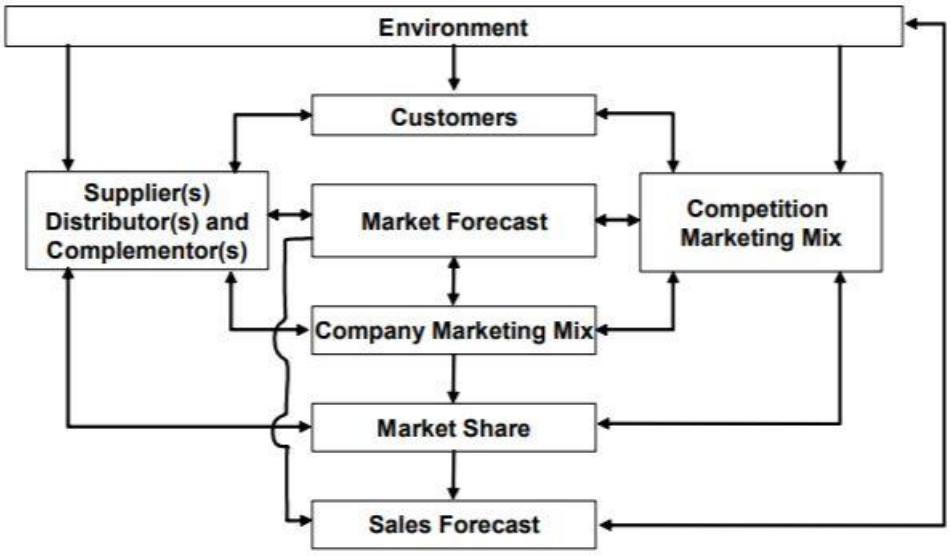


Figure 3 - Causal approach to sales forecasting

The primary advantage of the indirect approach is that it can be more directly related to decision making. Adjustments can be made in the marketing mix to see how this would

affect the forecast. Also, forecasts can be prepared to assess possible changes by other decision makers such as competitors or complementors. These forecasts can allow the firm to develop contingency plans, and these effects on 6 sales can also be forecast. On the negative side, the causal approach is more expensive than sales extrapolation.

FACTORS AFFECTING SALES:

Environment: It is sometimes possible to obtain published forecasts of environmental factors from Table base, which is available on the Internet through various subscribing business research libraries. These forecasts may be adequate for many purposes. However, sometimes it is difficult to determine what methods were used to create the forecasts. In such cases, econometric models can improve the accuracy of environmental forecasts. They provide more accurate forecasts than those provided by extrapolation or by judgment when large changes are involved. Allen (1999) summarizes evidence on this. Important findings that aid econometric methods are to:

- Base the selection of causal variables upon forecasting theory and knowledge about the situation, rather than upon the statistical fit to historical data (also, tests of statistical significance play no role here);
- Use relatively simple models (e.g. do not use simultaneous equations; do not use models that cannot be specified as linear in the parameters); and
- Use variables only if the estimated relationship to sales is in the same direction as specified a priori. The last point is consistent with the principle of using causal not statistical reasoning. Consistent with this viewpoint, leading indicators, a non causal approach to forecasting that has been widely accepted for decades, does not seem to improve the accuracy of forecasts (Diebold and Rudebusch 1991).

Interestingly, there exists little evidence that more accurate forecasts of the environment (e.g. population, the economy, social trends, technological change) lead to better sales forecasts. This, of course, seems preposterous. I expect that the results have been

obtained for studies where the conditions were not ideal for econometric methods. For example, if things continue to change as they have in the past, there is little reason to expect an econometric model to help with the forecast. However, improved environmental forecasts are expected when large changes are likely, such as the adoption of free trade policies, reductions in tariffs, economic depressions, natural disasters, and wars.

Customers: One should know the size of the potential market for the given product category (e.g. how many people in region X might be able to purchase an automobile), the ability of the potential market to purchase (e.g. income per capita and the price of the product), and the needs of the potential customers. Examination of each of these factors can help in forecasting demand for the category.

Company: The company sets its own marketing mix so there is typically little need to forecast these actions. However, sometimes the policies are not implemented according to plan because of changes in the market, actions by competitors or by retailers, or a lack of cooperation by those in the firm. Thus, it may be useful to forecast the actions that will actually be taken (e.g. if we provide a trade discount, how will this affect the average price paid by final consumers?)

Intermediaries: What actions will be taken by suppliers, distributors and complementors? One useful prediction model is to assume that their future decisions will be similar to those in the past, that is, the naive model. For existing markets, this model is often difficult to improve upon. When large changes are expected, however, the naive model is not appropriate. In such cases one can use structured judgment, extrapolate from analogous situations, or use econometric models.

Structure typically improves the accuracy of judgment, especially if it can realistically mirror the actual situation. Role playing is one such structured technique. It is useful when the outcome depends on the interaction among different parties and especially when the interaction involves conflict. Armstrong and Hutcherson (1989) asked subjects to role play the interactions between producers and distributors. In this disguised

situation, Philco was trying to convince supermarkets to sell its appliances through a scheme whereby customers received discounts based on the volume of purchases at selected supermarkets. Short (less than one hour) role plays of the situation led to correct predictions of the supermarket managers" responses for 75 per cent of the 12 groups. In contrast, only one of 37 groups was correct when groups made predictions without benefit of formal techniques. (As it turned out, the decision itself was poor, but that is another story.)

Econometric models offer an alternative, although much more expensive approach to forecasting the actions by intermediaries. This approach requires a substantial amount of information. For example, Montgomery (1975) described a model to predict whether a supermarket buying committee would put a new product on its shelves. This model, which used information about advertising, suppliers" reputation, margin and retail price, provided reasonable predictions for a hold-out sample.

Competitors: Can we improve upon the simple, "naïve," forecast that competitors will continue to act as they have in the past? These forecasts are difficult because of the interaction that occurs among the key actors in the market. Because competitors have conflicting interests, they are unlikely to respond truthfully to an intentions survey.

A small survey of marketing experts suggested that the most popular approach to forecasting competitors' actions is unaided expert opinion (Armstrong et al. 1987). Because the, experts' are usually those in the company, however, this may introduce biases related to their desired outcomes. For example, brand managers are generally too optimistic about their brands. Here again, role playing would appear to be relevant. Although no direct experimental evidence is available on its value in forecasting competitor's actions, role playing has proven to be accurate in forecasting the decision made in conflict situations (Armstrong 1999).

Market share: Can we do better than the naive model of no change? For existing markets that are not undergoing major change, the naive model is reasonably accurate (Brodie et al. 1999). This is true even when one has excellent data about the competitors

(Alsem et al. 1989). However, causal models should improve forecasts when large changes are made, such as when price reductions are advertised. Causal models should also help when a firm's sales have been artificially limited due to production capacity, tariffs, or quotas. Furthermore, contingent forecasts are important. Firms can benefit by obtaining good forecasts of how its policies (e.g. a major price reduction) would affect its market share.

CHAPTER -4

NEW PRODUCT FORECASTING

New product forecasting is of particular interest in view of its importance to decision making. In addition, large errors are typically made in such forecasts. Tull (1967) estimated the mean absolute percentage error for new product sales to be about 65 per cent. Not surprisingly then, pretest market models have gained wide acceptance among business firms; Shocker and Hall (1986) provide an evaluation of some of these models.

The choice of a forecasting model to estimate customer response depends on the stage of the product life-cycle. As one moves through the concept phase to the prototype, test market, introductory, growth, maturation, and declining stages, the relative value of the alternative forecasting methods changes. In general, the movement is from purely judgmental approaches to quantitative models that use judgment as inputs. For example, intentions and expert opinions are vital in the concept and prototype stages. Later, expert judgment is useful as an input to quantitative models. Extrapolation methods may be useful in the early stages if it is possible to find analogous products (Claycamp and Liddy 1969). In later stages, extrapolation methods become more useful and less expensive as one can work directly with time-series data on sales or orders. Econometric and segmentation methods become more useful after a sufficient amount of actual sales data are obtained.

When the new product is in the concept phase, a heavy reliance is usually placed on intentions surveys. Intentions to purchase new products are complicated because potential customers may not be sufficiently familiar with the proposed product and because the various features of the product affect one another (e.g. price, quality, and distribution channel). This suggests the need to prepare a good description of the proposed product. This often involves expensive prototypes, visual aids, product clinics,

or laboratory tests. However, brief descriptions are sometimes as accurate as elaborate descriptions as found in Armstrong and Overton's (1970) study of a new form of urban mass transportation.

In the typical intentions study, potential consumers are provided with a description of the product and the conditions of sale, and then are asked about their intentions to purchase. Eleven-point rating scales are recommended. The scale should have verbal designations such as 0 = No chance, almost no chance (1 in 100) to 10 = certain, practically certain (99 in 100). It is best to state the question broadly about one's "expectations" or "probabilities" to purchase, rather than the narrower question of intentions. This distinction was raised early on by Juster (1966) and its importance has been shown in empirical studies by Day et al. (1991).

Intentions surveys are useful when all of the following conditions hold:

- The event is important;
- Responses can be obtained;
- The respondent has a plan;
- The respondent reports correctly;
- The respondent can fulfill the plan; and
- Events are unlikely to change the plan.

These conditions imply that intentions are more useful for short-term forecasts of business-to-business sales.

The technology of intentions surveys has improved greatly over the past half century. Useful methods have been developed for selecting samples, compensating for non-Response bias and reducing response error. Dillman (1978) provides excellent advice that can be used for designing intentions surveys. Improvements in this technology have been demonstrated by studies on voter intentions (Perry 1979). Response error is probably the most important component of total error (Sudman and Bradburn 1982). Still, the correspondence between intentions and sales is often not close. Morwitz (1999) provides a review of the evidence on intentions to purchase.

As an alternative to asking potential customers about their intentions to purchase, one can ask experts to predict how consumers will respond. For example, Wotruba and Thurlow (1976) discuss how opinions from members of the sales force can be used to forecast sales. One could ask distributors or marketing executives to make sales forecasts. Expert opinions studies differ from intentions surveys. When an expert is asked to predict the behavior of a market, there is no need to claim that this is a representative expert. Quite the contrary, the expert may be exceptional. When using experts to forecast, one needs few experts, typically only between five and twenty (Hogarth 19, 78; Ashton 1985).

Experts are especially useful at diagnosing the current situation, which we might call "now casting." Surprisingly, however, when the task involves forecasting change, experts with modest domain expertise 9 (about the item to be forecast) are just as accurate as those with high expertise (Armstrong 1985: 91-6 reviews the evidence). This means that it is not necessary to purchase expensive expert advice.

Unfortunately, experts are often subject to biases. Salespeople may try to forecast on the low side if the forecasts will be used to set quotas. Marketing executives may forecast high in their belief that this will motivate the sales force. If possible, avoid experts who would have obvious reasons to be biased (Tyebjee 1987). Another strategy is to include a heterogeneous group of experts in the hopes that their differing biases may cancel one another.

Little is known about the relative accuracy of expert opinions versus consumer intentions. However, Sewall (1981) found that each approach contributes useful information such that a combined forecast is more accurate than either one alone.

Producers often consider several alternative designs for the new product. In such cases, potential customers can be presented with a series of perhaps twenty or so alternative offerings. For example, various features of a personal computer, such as price, weight, battery life, screen clarity and memory might vary according to rules for experimental design (the basic ideas being that each feature should vary substantially and that the

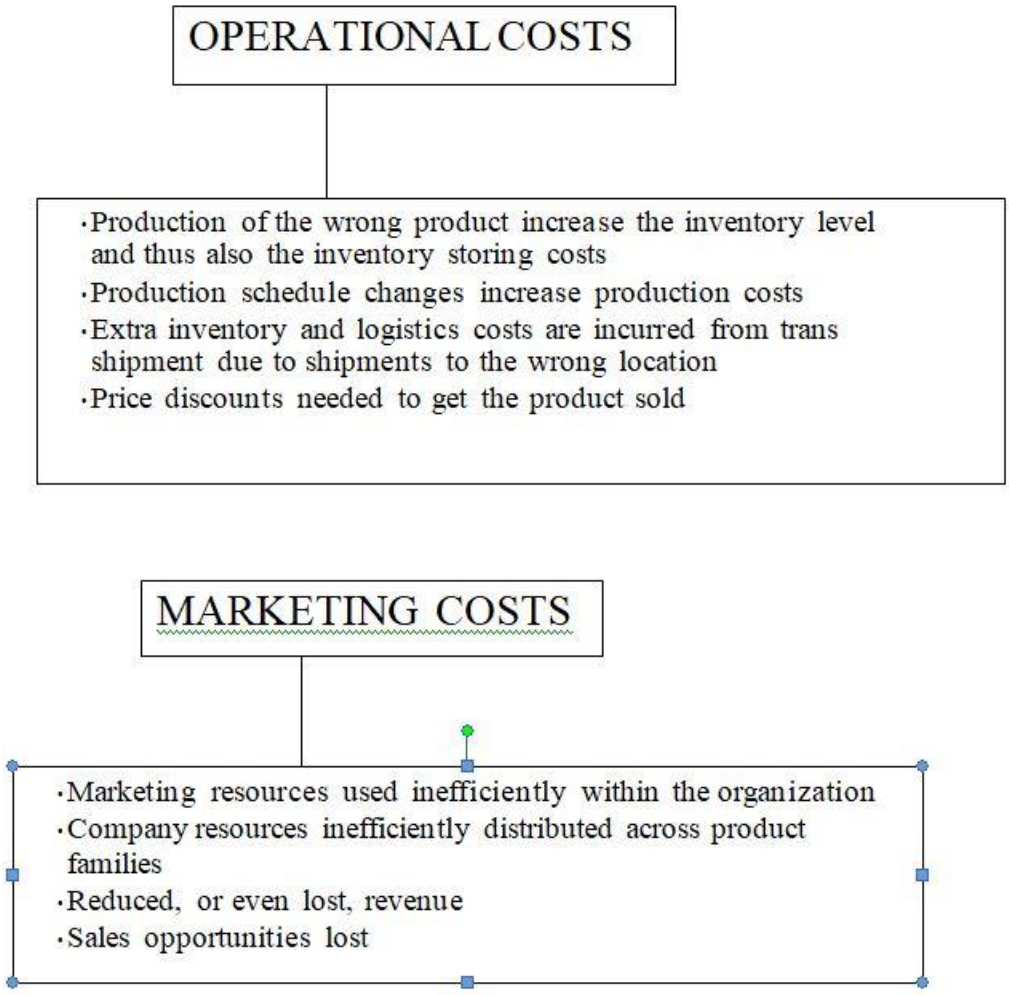
variations among the features should not correlate with one another). The customer is forced to make trade-offs among various features. This is called "conjoint analysis" because the consumers consider the product features jointly. This procedure is widely used by firms (Wittink and Bergestuen 1998). An example of a successful application is the design of a new Marriott hotel chain (Wind et al. 1989). The use of conjoint analysis to forecast new product demand can be expensive because it requires large samples of potential buyers, the potential buyers may be difficult to locate, and the questionnaires are not easy to complete. Respondents must, of course, understand the concepts that they are being asked to evaluate. Although conjoint analysis rests on good theoretical foundations, little validation research exists in which its accuracy is compared with the accuracy of alternative techniques such as Delphi or judgmental forecasting procedures.

Expert judgments can be used in a manner analogous to the use of consumers' intentions for conjoint analysis. That is, the experts could be asked to make predictions about situations involving alternative product design and alternative marketing plans. These predictions would then be related to the situations by regression analysis. Following the philosophy for naming conjoint analysis, this could be called exjoint analysis. It is advantageous to conjoint analysis in that few experts are needed (probably between five and twenty). In addition, it can incorporate policy variables that might be difficult for consumers to assess.

Once a new product is on the market, it is possible to use extrapolation methods. Much attention has been given to the selection of the proper functional form to extrapolate early sales. The diffusion literature uses an S-shaped curve to predict new product sales. That is, growth builds up slowly at first, becomes rapid as word-of-mouth and observation of use spread, then slows again as it approaches a saturation level. A substantial literature exists on diffusion models. Despite this, the number of comparative validation studies is small and the benefits of choosing the best functional form seem to be modest (research on this is reviewed by Meade 1999).

Effects Of An Error: Although the accuracy of forecasts is known within

an organization, the financial impact of an error in it might not be as apparent. Kahn (2003) describes in his article how a forecasting error has impact on an organization. The method described derives an approximate figure, though; it still gives a good picture of what the financial impact of an error can be. He identifies costs related to a forecast error and separates them into



operational costs and marketing costs. These different costs are related to the forecast error and the variations of the two types can be incurred by two different scenarios; an over-forecast and an under-forecast. When the organization plans its operations from an over-forecast, extra cost will incur. Extra costs would be incurred if the organization would have

chosen to base their operations on an under-forecast although different extra costs incurs in the two situations. (Mentzer, 1999)

To reduce the forecast error, several generalizations of the more successful methods can be drawn. (Armstrong, 2006) The first generalization argues that a forecaster needs to be conservative when uncertain, in order to reduce forecast error. Further on, the need to spread the risk is argued. By decompose, segment and combine methods a forecaster spread the risk compared to if only one method is being used. Another important aspect to reduce the error is the use of realistic representations of the situation. Methods that use more information are generally more accurate than methods using only one source of information. Furthermore, methods relying only on data are inferior to methods using prior knowledge about relationships and situations. The last generalization to reduce forecast error is that structured methods are generally more accurate than unstructured. These generalizations are helpful for a company when deciding what forecasting method to use.

Errors associated with a survey can be grouped within two main groups; random sampling error and systematic errors. (Zikmund, 2003) In our case, the systematic errors were the ones focused on to minimize since our respondents were not randomly chosen. Within the systematic error category, there are two broad groups where errors

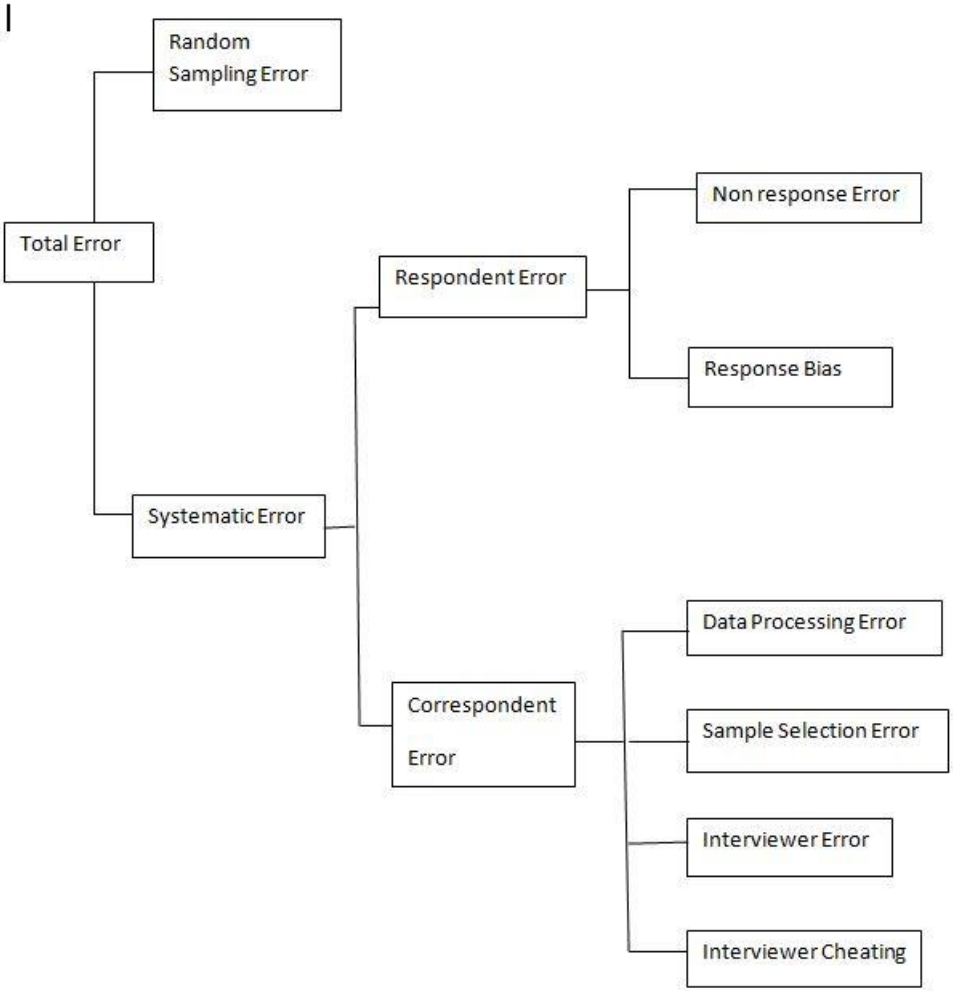
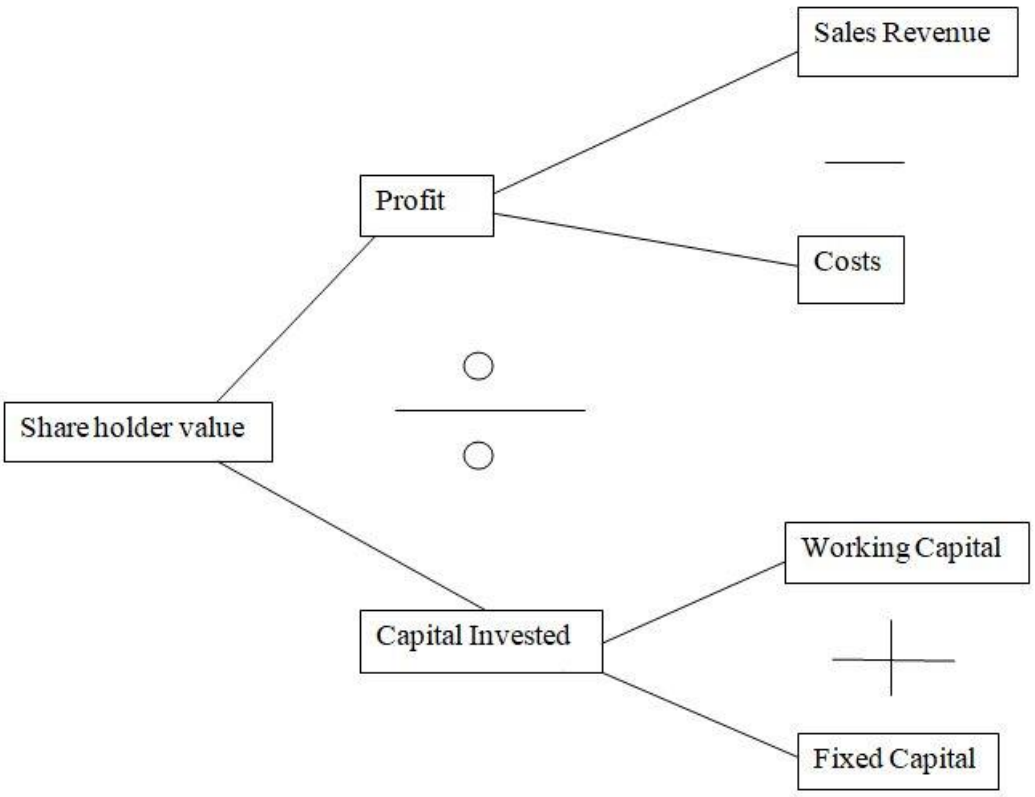


Figure 4 – Different errors in surveying

Impact return on share holder value: As a consequence of an inaccurate forecast, the organization will induce extra costs,which in the end will have a negative impact on the return to shareholder's value. (Mentzer, 1999) By having an accurate forecast and therefore managing the inventory Level,the organization can reduce the inventory carrying costs and also decrease the trans shipment costs . A seasonal accurate forecast could also increase the sales

revenue due to the fact that stock will not sell out or that the inventory will not grow too large. Buy reducing the costs and increasing the revenue.

the revenue, the profit will be greater. The other side of the return to shareholder's value is the rate of capital turnover. If the forecast is accurate, less investment in inventories is needed and therefore the capital invested decreases. An increased profit together with a decreased invested capital sums up in an increase in the return to shareholder's value. The relationships are shown in figure 4 between the different components.



CHAPTER -5

PROPOSED MODEL

Multiple Linear Regression: It is a statistical model which shows the relationship between two or more variables one being the dependent variable and other being the independent variable with a linear equation. Here we are using linear regression to find out the relation between the sales and other factors that can affect the sales and we check how better this model helps in determining the relation between sales and the attributes which affect them.

$$Y = b_1X_1 + b_2X_2 + \ldots. + b_nX_n$$

The above formula represents the relation between various independent variables i.e. attributes with that to the dependent variable i.e. sales.

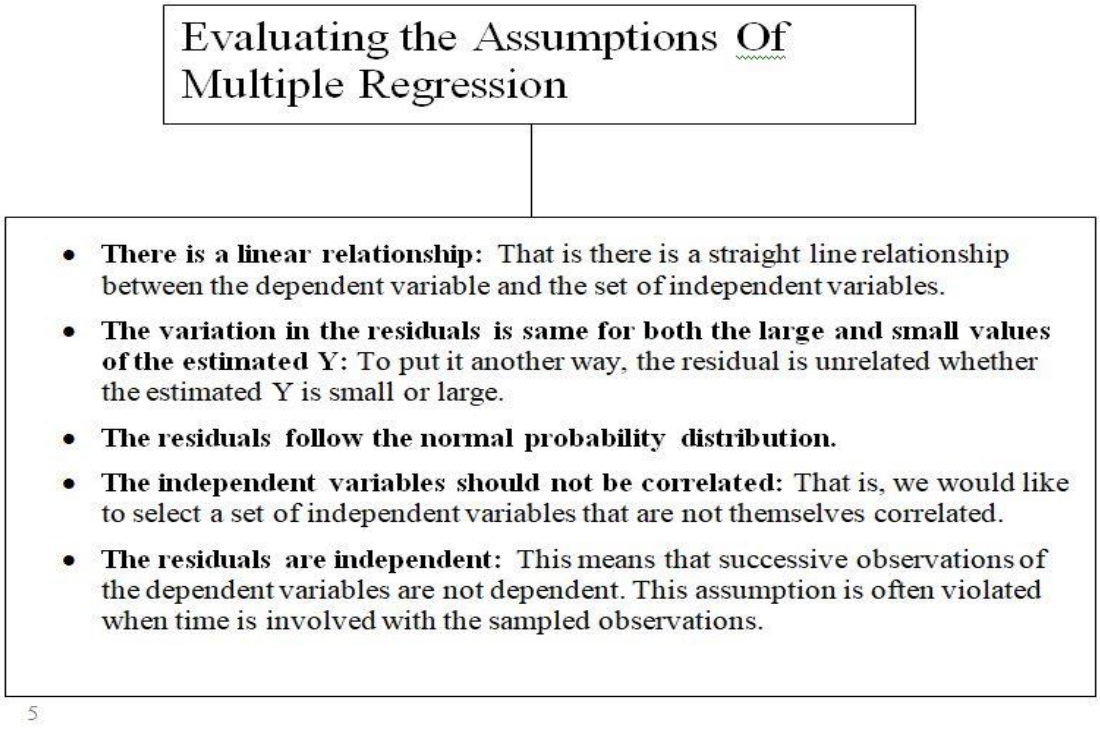


Figure 5 – Multiple linear regressions

Decision Tree: This is an algorithm which uses classification to determine the various conditions in order to take a decision. Here we use this tree to determine the sales of a certain product depending on a single attribute. These trees help in determining the effect of certain attribute on the sales of that product.

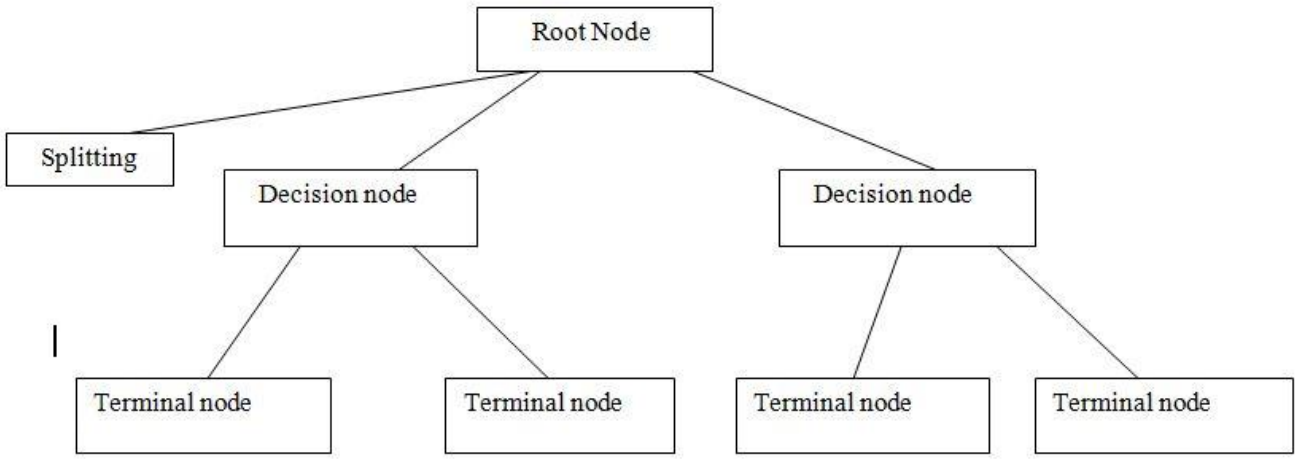


Fig. Decision tree

Figure 6 - Node information in decision trees

Ridge regression: Ridge regression uses the l2 regularization which allows to create a model when the number of predictor variables in a set exceeds the no. of observations. It is able to work with multi collinear data. It does not face the problem of over fitting. Here the penalty is on the sum of squared coefficients.

Validation Score (RMSE) -
1131.79
Leader board Score (RMSE) –
1203.56

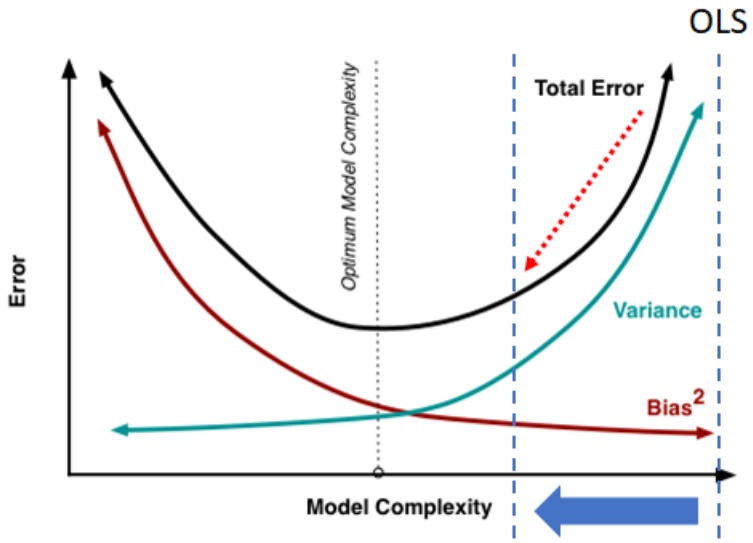


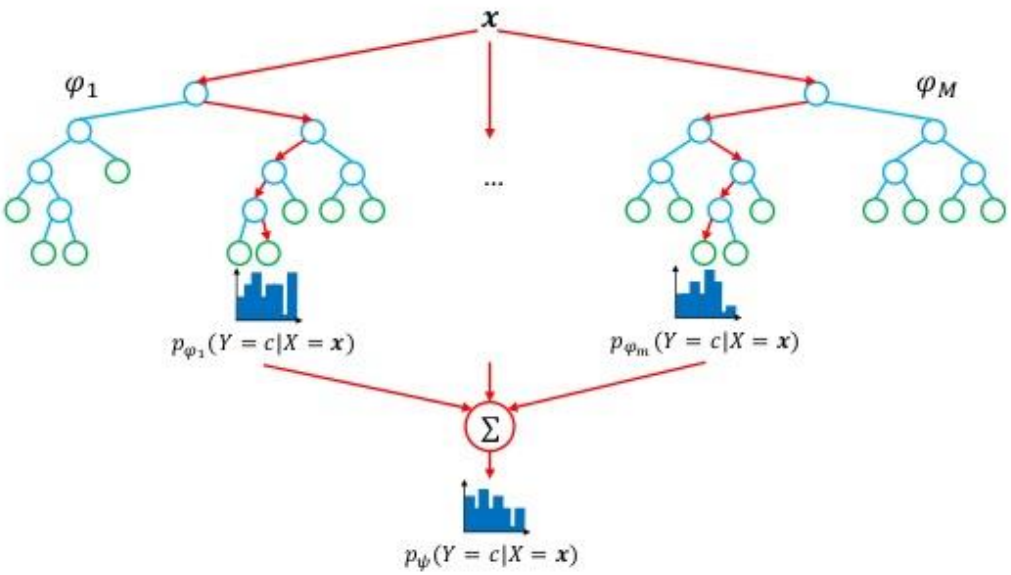
Figure 7 – Ridge regression graph

Lasso regression: It is a regression analysis method which performs the task of variable selection and also uses the l2 regularization. Here we use a penalty which changes the value of coefficients of regression. Here the penalty is on the sum of absolute values of coefficients which helps in removing over fitting.

Validation Score (RMSE) –
1127.45
Leader board Score (RMSE) –
1200.67

Random forests: It is a higher version of decision trees , in this method the data set is divided into various samples and the decision tree algorithm is applied to each sample separately during the training phase. The results which come from each sample and then a majority vote is taken out in order to find the final result .It also removes the overfitting problem of decision trees. It improves the accuracy as compared to decision trees and also estimates the missing values.

Random forests



Randomization

- Bootstrap samples
- Random selection of $K \leqslant p$ split variables } Random Forests
- Random selection of the threshold } Extra-Trees

14 / 39

Figure 8 - Random forest explanation

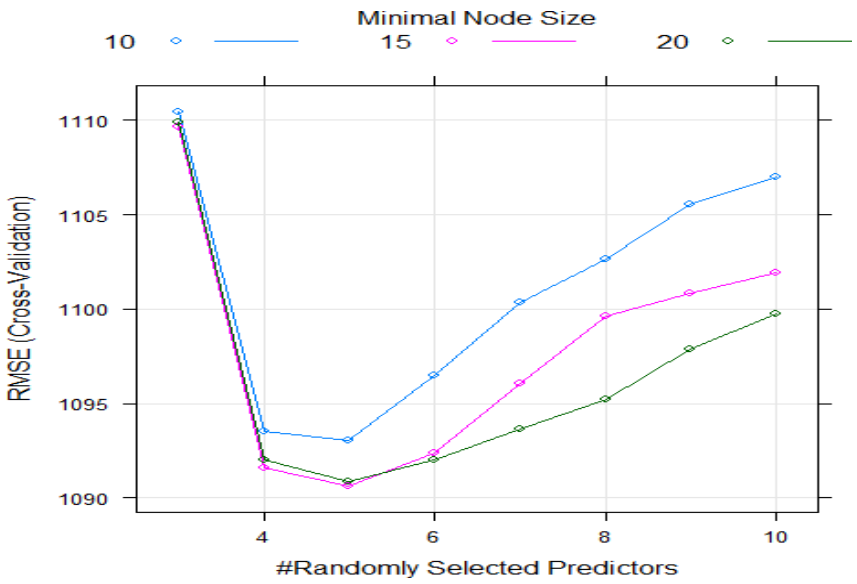


Figure 9 - Best model parameters

Validation Score (RMSE) – 1094.23 Leader board Score (RMSE) – 1147.89

XGBoost: It is the extreme gradient boosting method. It is an ensemble technique based on boosting. It is a sequential decision tree based learning algorithm .here we assign weights to different samples of the dataset and each sample is analyzed by decision trees and weak classifiers are generated and keep increasing with each classifier and ultimately gives better results but it is a bit regressive and bit more time taking as other algorithms.

Validation Score (RMSE) –
1085.11
Leader board Score (RMSE) –
1142.31

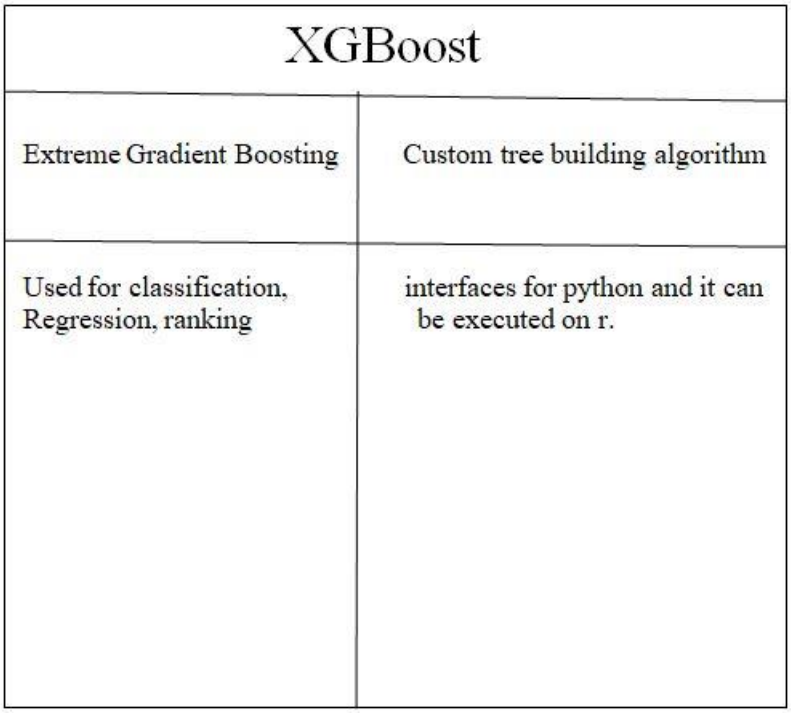


Figure 10 - XGboost Explanation

CHAPTER-6

IMPLEMENTATION OR ARCHITECTURAL DESIGN

Module Split-Up

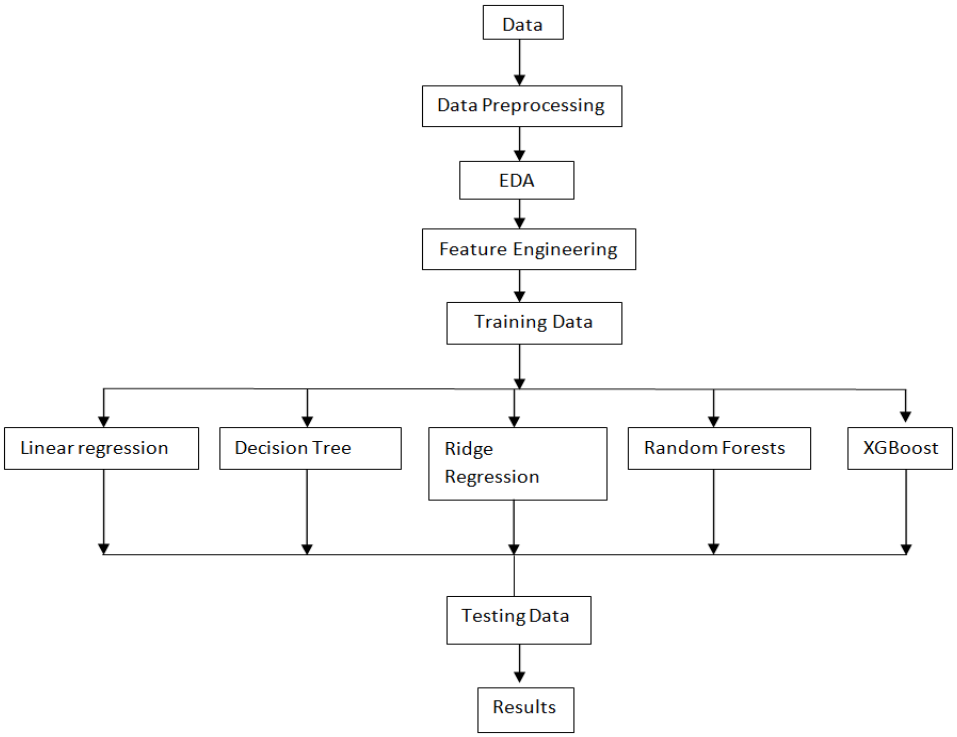


Figure 11- Flowchart for sales forecasting via different techniques

The following flowchart describes the workflow of this paper

STEP 1: Loading the data into the environment

STEP 2: Data preprocessing.

STEP 3: Performing feature selection and subsetting important features.

STEP 4: Splitting the data into train and test sets.

STEP 5: Model the training data and form different classifiers.

STEP 6: Pass the test data to all trained classifiers.

STEP 7: Evaluate using a confusion matrix and measure accuracy and precision

IMPLEMENTATION :

Step 1: DATA EXPLORATION: Exploring data from various data sources and importing data to research on it.

Import data

```
import pandas as pd

import numpy as np



#Read files:

train = pd.read_csv("train.csv")

test = pd.read_csv("test.csv")
```

Training and Testing Data

```
import pandas as pd

import numpy as np



#Read files:

train = pd.read_csv("train.csv")

test = pd.read_csv("test.csv")
```

Describing Data:

```
data.apply(lambda x: sum(x.isnull()))
```

```
data.describe()
```

```
data.apply(lambda x: len(x.unique()))
```

Filter Categorical Variables:

```
categorical_columns = [x for x in data.dtypes.index if
data.dtypes[x]=='object']
```

Exclude ID columns and groups:

```
categorical_columns = [x for x in categorical_columns if x not in
['Item_Identifier','Outlet_Identifier','source']]
```

Print frequencies of categories:

```
print '\nFrequency of Categories for varible %s'%col

    print data[col].value_counts()
```

Step 2 -DATA CLEANING: It is the process of cleaning data and preparing it in the form such that it does not contain any kind of noises or unstructured data

Determining average weight of each item:

```
item_avg_weight = data.pivot_table(values='Item_Weight',
index='Item_Identifier')
```

Get a boolean variable specifying various weights:

```
miss_bool = data['Item_Weight'].isnull()
```

Impute data and check missing values before and after imputing:

```
print 'Orignal #missing: %d'% sum(miss_bool)

data.loc[miss_bool,'Item_Weight'] =
data.loc[miss_bool,'Item_Identifier'].apply(lambda x:
item_avg_weight[x])

print 'Final #missing: %d'% sum(data['Item_Weight'].isnull())
```

Import Mode Function:

```
from scipy.stats import mode
```

Determining the mode of each:

```
outlet_size_mode = data.pivot_table(values='Outlet_Size',
columns='Outlet_Type',aggfunc=(lambda x:mode(x).mode[0]) )

print 'Mode for each Outlet_Type:'

print outlet_size_mode
```

Get a Boolean Variable Specifying Item_ Weights Values:

```
miss_bool = data['Outlet_Size'].isnull()
```

Impute data and check missing values before and after imputation to confirm

```
print '\nOrignal #missing: %d'% sum(miss_bool)

data.loc[miss_bool,'Outlet_Size'] =
data.loc[miss_bool,'Outlet_Type'].apply(lambda x:
outlet_size_mode[x])

print sum(data['Outlet_Size'].isnull())
```

Step 4 – FEATURE ENGINEERING:

It refers to dealing with nuances that is present in data and making our data ready for analysis.

Combining Outlet_type:

```
data.pivot_table(values='Item_Outlet_Sales',index='Outlet_Type')
```

Modify Item_Visibility:

```
visibility_avg = data.pivot_table(values='Item_Visibility',
index='Item_Identifier')

miss_bool = (data['Item_Visibility'] == 0)

print 'Number of 0 values initially: %d'%sum(miss_bool)

data.loc[miss_bool,'Item_Visibility'] =
data.loc[miss_bool,'Item_Identifier'].apply(lambda x:
visibility_avg[x])

print 'Number of 0 values after modification:
%d'%sum(data['Item_Visibility'] == 0)

data['Item_Visibility_MeanRatio'] = data.apply(lambda x:
x['Item_Visibility']/visibility_avg[x['Item_Identifier']], axis=1)

print data['Item_Visibility_MeanRatio'].describe()
```

Create a broad category of type of item:

```
data['Item_Type_Combined'] = data['Item_Identifier'].apply(lambda x:
x[0:2])

data['Item_Type_Combined'] =
data['Item_Type_Combined'].map({'FD':'Food',

                                'NC':'Non-
Consumable',
```

```
'DR':'Drinks'})

data['Item_Type_Combined'].value_counts()
```

Determining the years of operation of store:

```
#Years: data.loc[data['Item_Type_Combined']=="Non-
Consumable",'Item_Fat_Content'] = "Non-Edible"

data['Item_Fat_Content'].value_counts()



data['Outlet_Years'] = 2013 - data['Outlet_Establishment_Year']

data['Outlet_Years'].describe()
```

Modify Category of item_fat_Content:

```
print 'Original Categories:'

print data['Item_Fat_Content'].value_counts()



print '\nModified Categories:'
```

```
data['Item_Fat_Content'] = data['Item_Fat_Content'].replace({'LF':'Low
Fat',


'reg':'Regular',

                                                          'low
fat':'Low Fat'})

print data['Item_Fat_Content'].value_counts()
```

Numerical and one hot encoding of Categorical variables:

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

data['Outlet'] = le.fit_transform(data['Outlet_Identifier'])

var_mod =
['Item_Fat_Content','Outlet_Location_Type','Outlet_Size','Item_Type_Com
bined','Outlet_Type','Outlet']

le = LabelEncoder()

for i in var_mod:
```

```
    data[i] = le.fit_transform(data[i])

data = pd.get_dummies(data,
columns=['Item_Fat_Content','Outlet_Location_Type','Outlet_Size','Outle
t_Type',

                              'Item_Type_Combined','Outlet'])

data.dtypes

data[['Item_Fat_Content_0','Item_Fat_Content_1','Item_Fat_Content_2']].
head(10)
```

Exporting Data:

```
data.drop(['Item_Type','Outlet_Establishment_Year'],axis=1,inplace=True
)

train = data.loc[data['source']=="train"]

test = data.loc[data['source']=="test"]

test.drop(['Item_Outlet_Sales','source'],axis=1,inplace=True)

train.drop(['source'],axis=1,inplace=True)

train.to_csv("train_modified.csv",index=False)
```

```
test.to_csv("test_modified.csv",index=False)
```

Step 4 – Model Building:

```
mean_sales = train['Item_Outlet_Sales'].mean()

base1 = test[['Item_Identifier','Outlet_Identifier']]

base1['Item_Outlet_Sales'] = mean_sales

base1.to_csv("alg0.csv",index=False)

target = 'Item_Outlet_Sales'

IDcol = ['Item_Identifier','Outlet_Identifier']

from sklearn import cross_validation, metrics

def modelfit(alg, dtrain, dtest, predictors, target, IDcol, filename):
```

Checking Each Model:

Linear Regression Model:

```
from sklearn.linear_model import LinearRegression, Ridge, Lasso
```

```
predictors = [x for x in train.columns if x not in [target]+IDcol]

alg1 = LinearRegression(normalize=True)

modelfit(alg1, train, test, predictors, target, IDcol, 'alg1.csv')

coef1 = pd.Series(alg1.coef_, predictors).sort_values()

coef1.plot(kind='bar', title='Model Coefficients')
```

Ridge Regression Model:

```
predictors = [x for x in train.columns if x not in [target]+IDcol]

alg2 = Ridge(alpha=0.05,normalize=True)

modelfit(alg2, train, test, predictors, target, IDcol, 'alg2.csv')

coef2 = pd.Series(alg2.coef_, predictors).sort_values()

coef2.plot(kind='bar', title='Model Coefficients')
```

Decision Tree Model:

```
from sklearn.tree import DecisionTreeRegressor
```

```
predictors = [x for x in train.columns if x not in [target]+IDcol]

alg3 = DecisionTreeRegressor(max_depth=15, min_samples_leaf=100)

modelfit(alg3, train, test, predictors, target, IDcol, 'alg3.csv')

coef3 = pd.Series(alg3.feature_importances_,
predictors).sort_values(ascending=False)

coef3.plot(kind='bar', title='Feature Importances')

predictors = ['Item_MRP','Outlet_Type_0','Outlet_5','Outlet_Years']

alg4 = DecisionTreeRegressor(max_depth=8, min_samples_leaf=150)

modelfit(alg4, train, test, predictors, target, IDcol, 'alg4.csv')

coef4 = pd.Series(alg4.feature_importances_,
predictors).sort_values(ascending=False)

coef4.plot(kind='bar', title='Feature Importances')
```

Random Forest Model:

```
from sklearn.ensemble import RandomForestRegressor

predictors = [x for x in train.columns if x not in [target]+IDcol]
```

```
alg5 = RandomForestRegressor(n_estimators=200,max_depth=5,
min_samples_leaf=100,n_jobs=4)

modelfit(alg5, train, test, predictors, target, IDcol, 'alg5.csv')

coef5 = pd.Series(alg5.feature_importances_,
predictors).sort_values(ascending=False)

coef5.plot(kind='bar', title='Feature Importances')

predictors = [x for x in train.columns if x not in [target]+IDcol]

alg6 = RandomForestRegressor(n_estimators=400,max_depth=6,
min_samples_leaf=100,n_jobs=4)

modelfit(alg6, train, test, predictors, target, IDcol, 'alg6.csv')

coef6 = pd.Series(alg6.feature_importances_,
predictors).sort_values(ascending=False)

coef6.plot(kind='bar', title='Feature Importances')
```

CHAPTER-7

RESULTS AND DISCUSSION

The experiment is carried out on a computer with the Windows 10 operating system, 16GB of RAM, and 1TB or hard drive.

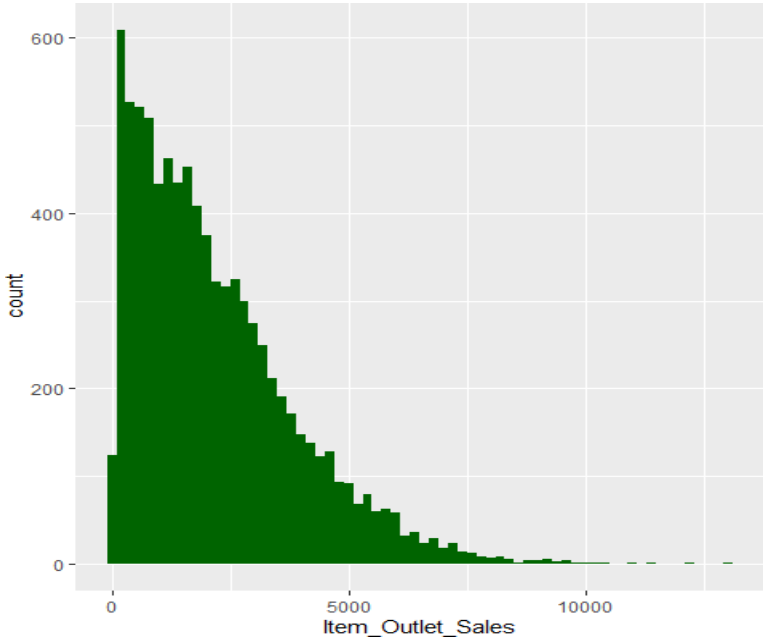


Figure 12 - Target Variable

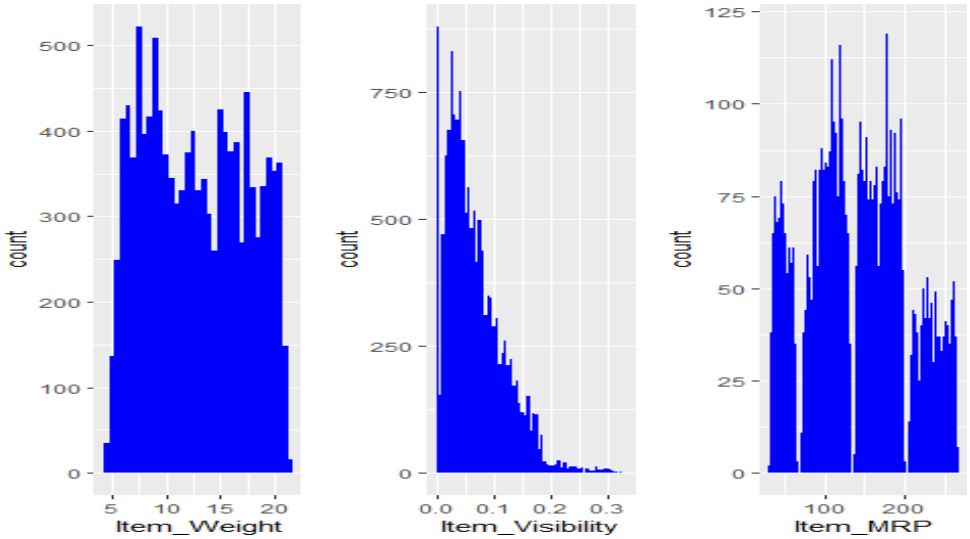


Figure 13 - Uni variate Analysis 1

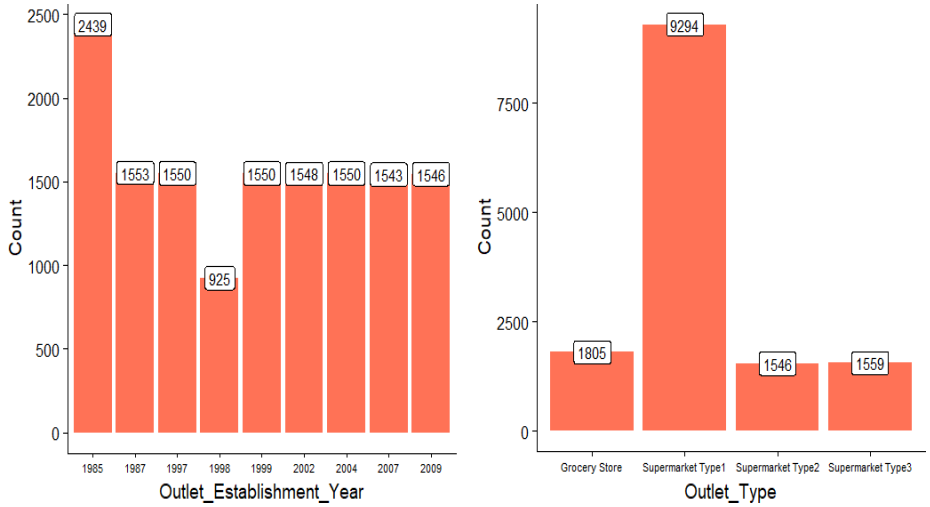


Figure 14 - Uni variate analysis 2

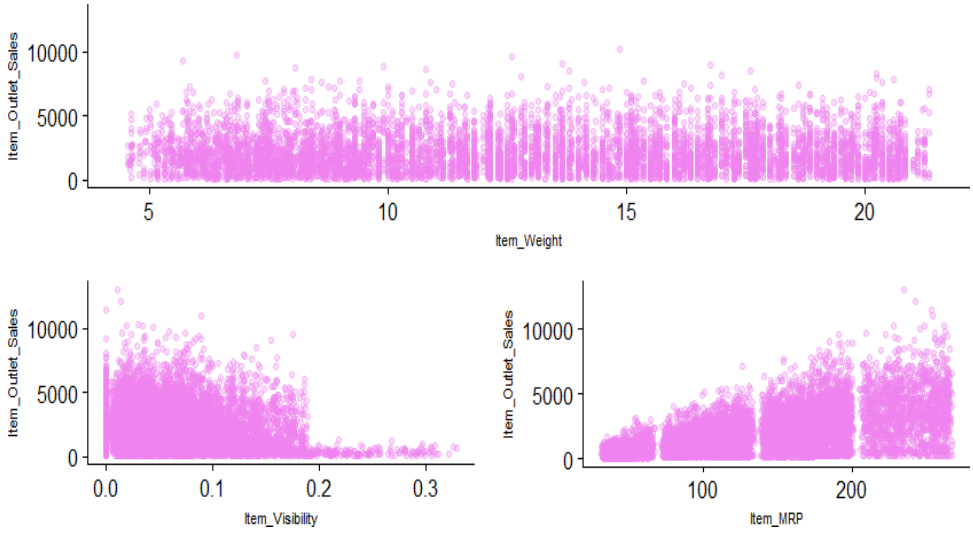


Figure 15 - Uni variate analysis 3

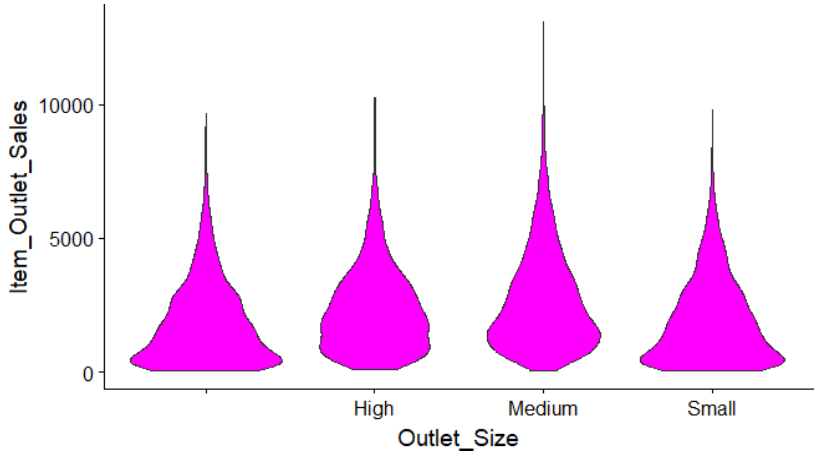


Figure 16 - Uni variate Analysis 4

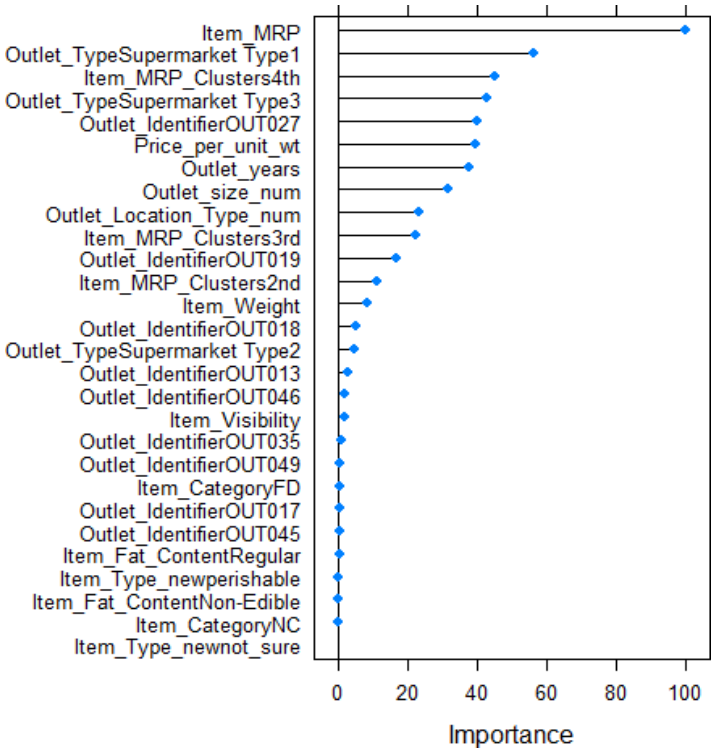


Figure 17 - Variable Importance

The following shows the accuracy of all the models when used on the test dataset.

Algorithm	Cross Validation Score	RMSE Score
Linear Regression	1156.32	1199.34
Ridge Regression	1131.79	1203.56
Lasso Regression	1127.45	1200.67
Random Forest	1094.23	1147.89
XGBoost	1085.11	1142.31

Figure 18 - Different algorithm performance

From the above data, it is evident that XGboost is the best-fit algorithm for this classification algorithm as suggested in the proposed model. However, it was also the model that took the longest time to execute so we might also need to consider the time- accuracy trade-off when dealing with this situation in real-time.

CHAPTER-8

CONCLUSIONS AND FUTURE WORKS

This paper shows us the different techniques that are used in order to identify that what is the model that is required by different companies according to the accuracy-cost trade off . Based on the above result we could see that XGboost turned out to be the best working model for this problem in terms of accuracy as it takes multi collinearity in account while random forests doesn't account for it but the cost of performing XGboost is way higher as compared to random forests. So it basically depends on the company whether it opts for accuracy or to reduce its cost and its goal is to minimize the cost function which makes it an optimized or near optimized solution. The results from the decision tree and random forest were accurate too because of the information gain by entropy which takes place on every split made by a tree and hence giving more information to the machine. Multiple linear regression, lasso regression, ridge regression weren't as accurate as of the other because it just depends on just one variable at a time while random forests and XGboost take all trees into account. Lastly, we conclude that based on the market analysis we concluded find confidence among several products which would be fruitful when the retailer is able to organize its store in an efficient manner for its increase in the sale. Thus, this paper draws some conclusions on how the customer behaviors can be judged and anticipated beforehand so that necessary changes can be made to retain the customer.

Comparing the results of regression techniques and boosting techniques like XGBoost will be very close in terms of precision and recall. Hybrid models can also be built in order to increase the accuracy of the system. Then, their accuracies can be measured in the same manner as done in this paper.. This system would be beneficial for both sellers and buyers. Using the transactional data, an efficient recommendation system can be built and hence the customers with similar liking will be suggested products that are available in the store.

In all the final findings show that the higher the order of multi collinearity the better is the

algorithm which uses gradient boosting in the extreme manner .this also depends on the accuracy of forecasting which is needed.

CHAPTER-9

References

[1] H. M. Al-Hamadi “Long-Term Electric Power Load Forecasting Using Fuzzy Linear Regression Technique”, IEEE Mar.2011

[2] Yanming Yang “Prediction and Analysis of Aero-Material Consumption Based on Multivariate Linear Regression Model”, 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis

[3] You Li Feng, Shan Shan Wang “A Forecast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression”, IEEE Dec.2013.

[4] Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., L¨utjen, M., Teucke, M.: A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering 3(1), 154–161 (2015)

[5] Bose, I., Mahapatra, R.K.: Business data mining a machine learning perspective. Information & management 39(3), 211–225 (2001)

[6]Ackoff, R.L., 1970, A Concept of Corporate Planning (John Wiley and Sons, New York).

[7]Ansoff, H.I, 1964, A quasi-analytical approach of the busi- ness strategy problem, Management Technology, IV, 67- 77.

[8]Ansoff, H.I., 1966, Corporate Strategy (McGraw Hill, New York).

[9]Ansoff, H.I., 1979, Strategic Management (Macmillan, Lon- don).

[10]Ansoff, H.I., 1994, Comment on Henry Mintzberg's rethink- ing strategic

planning, Long Range Planning, 27(3), 31-32.

[11]Bain and Company and the Planning Forum, 1995, Manage- ment Tools and Techniques: An Executive's Guide, 1995 (Bain and Company, Boston).