**Data Analysis of Sales Prediction of Big Mart Data**



A project paper is submitted for the degree of B.Sc. Examination 2018 in Computer Science and Engineering

**Project supervisor**

Dr. Shamim Ahmad

Professor

Dept. of Computer Science and Engineering,

University of Rajshahi

**Submitted BY**

Name: Md. Ismail Hossain

ID: 1510276139

Session: 2014-15

Exam: B.Sc. part-4 even semester

Dept. of Computer Science and Engineering,

University of Rajshahi

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY OF RAJSHAHI

**CHAPTER ONE**

**INTRODUCTION**

* 1. **Introduction**

The objective of this project is to analyze sales information of some popular e-commerce sites like amazon, e-bay, Alibaba, Coursera and Udacity. This analysis will produce some report on their sales’ pattern, relation between sales’ rate and price of the product etc. This type report may help to make a good business planning. Here I have chosen BigMart data for analysis.

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales

* 1. **Why this project is being chosen**

This project will analyze the BigMart sales data and help to take better decision for future. This project will help to make better sales plan for future

* 1. **Objective**

The aim is to build a predictive model and find out the sales of each product at a particular store.

Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

* 1. **Organization of the project**

This project is organized in details in this project paper. The hypothesis generation and coding issue is illustrated in chapter 2. This chapter describe about the python 3.5, numpy and panda (Data centric python packages). In chapter 3 elaboration of analysis

1. Hypothesis generation
2. Data exploring
3. Data cleaning
4. Feature engineering
5. Model building
   1. **Conclusion**

The data analysis of online business site, social media etc. are drawing much attention of data scientists in recent days.

So it’s the time to learn data analysis for making better decision about the future.

**CHAPTER TWO**

**BACKGROUND STUDY**

**2.1 Introduction**

Data analysis is a process of inspecting, [cleansing](https://en.wikipedia.org/wiki/Data_cleansing), [transforming](https://en.wikipedia.org/wiki/Data_transformation), and [modeling](https://en.wikipedia.org/wiki/Data_modeling) [data](https://en.wikipedia.org/wiki/Data) with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a [process](https://en.wikipedia.org/wiki/Process_theory) for obtaining raw data and converting it into information useful for decision-making by users. Data are collected and analyzed to answer questions, test hypotheses or disprove theories

**2.2 Data analysis**

There are total five steps to analyze a data set

1. Hypothesis generation
2. Data exploring
3. Data cleaning
4. Feature engineering
5. Model building

**Hypothesis generation:**

A statistical hypothesis, sometimes called confirmatory data analysis, is a [hypothesis](https://en.wikipedia.org/wiki/Hypothesis) that is testable on the basis of [observing](https://en.wikipedia.org/wiki/Observable_variable) a process that is [modeled](https://en.wikipedia.org/wiki/Statistical_model) via a set of [random variables](https://en.wikipedia.org/wiki/Random_variable).A statistical hypothesis test is a method of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference). Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an [alternative](https://en.wikipedia.org/wiki/Alternative_hypothesis) to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed [statistically significant](https://en.wikipedia.org/wiki/Statistically_significant) if the relationship between the data sets would be an unlikely realization of the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) according to a threshold probability—the significance level. Hypothesis tests are used when determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

Understanding the problem better by brainstorming possible factors that can impact the outcome.

**Data exploring:**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, like python and R.

Before it can conduct [analysis](https://whatis.techtarget.com/definition/statistical-analysis) on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

Looking at categorical and continuous feature summaries and making inferences about the data.

**Data cleaning:**

**Data cleansing** or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate [records](https://en.wikipedia.org/wiki/Storage_record) from a record set, [table](https://en.wikipedia.org/wiki/Table_(database)), or [database](https://en.wikipedia.org/wiki/Database) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the [dirty](https://en.wikipedia.org/wiki/Dirty_data) or coarse data. Data cleansing may be performed [interactively](https://en.wikipedia.org/wiki/Interactively) with [data wrangling](https://en.wikipedia.org/wiki/Data_wrangling) tools, or as [batch scripting processing](https://en.wikipedia.org/wiki/Batch_processing).

So, cleaning is the part of imputing missing values in the data and checking for outliers.

**Feature Engineering:**

Feature engineering is the process of using [domain knowledge](https://en.wikipedia.org/wiki/Domain_knowledge) of the data to create [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) that make [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms work. Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive.

Feature engineering is an essential part of building any intelligent system. Even though you have a lot of newer methodologies coming in like deep learning and meta-heuristics which aid in automated machine learning, each problem is domain specific and better features (suited to the problem) is often the deciding factor of the performance of your system. Feature Engineering is an art as well as a science and this is the reason Data Scientists

often spend 70% of their time in the data preparation phase before modeling.

So, Feature Engineering is modifying existing variables and creating new ones for analysis

**Model Building:**

Machine learning happens to be a small part of this process. The model building process involves setting up ways of collecting data, understanding and paying attention to what is important in the data to answer the questions you are asking, finding a statistical, mathematical or a simulation model to gain understanding and make predictions.

All of these things are equally important and model building is a crucial skill to acquire in every field of science. The process stays true to the scientific method, making what you learn through your models useful for gaining an understanding of whatever you are investigating as well as make predictions that hold true to test.

**CHAPTER THREE**

**TOOLS**