VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum – 590 018

Internship Report on

"INVESTMENT AND SALES PREDICTION USING MACHINE LEARNING"

Submitted in partial fulfilment of the requirements of the **VIII Semester Degree** of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

For the Academic Year: 2022-2023

By

PREKSHA M [1DB19CS108]

Internship carried out at

INVENTERON TECHNOLOGIES AND BUSINESS SOLUTION LLP

Internal Guide

Prof. Komala D

Asst.Prof., Dept. of CSE,DBIT,

Bangalore

External Guide

Mr. Syed Azad

Managing Director

Inventeron Technologies

DON BOSCO INSTITUTE OF TECHNOLOGY

Kumbalagodu, Bangalore – 560074

DECLARATION

I PREKSHA M, student of eighth-semester B.E, Department of Computer Science and Engineering, Don Bosco Institute of Technology, Kumbalagodu, Bengaluru, declare, that the internship work entitled –INVESTMENT AND SALES PREDICTION USING MACHINE LEARNING‖ has been carried out by me and submitted in partial fulfilment of the course requirements for the award of degree in Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belgaum during the academic year 2022-2023. The matter embodied in this report has not been submitted to any university or institute for the award of any other degree or diploma.

Place: Bangalore

Date:

PREKSHA M
1DB19CS108

DON BOSCO INSTITUTE OF TECHNOLOGY

Kumbalagodu, Bangalore – 560074

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the internship report entitled ‑INVESTMENT AND SALES PREDICTION USING MACHINE LEARNING‖ is a work carried out by PREKSHA M [1DB19CS108] in partial fulfilment of the award of Degree of Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi, during the academic year 2022-2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated. The internship report has been approved as it satisfies the academic requirements associated with the degree mentioned.

Signature of the Guide

Prof . Komala D
Asst.Prof., Dept. of CSE,
DBIT, Bangalore

Signature of the HOD

Dr. K B Shiva Kumar
Dept. of CSE,
DBIT, Bangalore

Principal

Dr. B.S. Nagabhushana
Principal,
DBIT, Bangalore

Name of External Examinear.

1._____

2._____

Signature with date

1. _____

2. _____

ABSTRACT

Predicting Investment and sales of a company needs time series data of that company and based on that data the model can predict the future sales of that company or product. So, in this research project we will analyse the time series sales data of a company and will predict the sales of the company for the coming quarter and for a specific product. For this kind of project of sales predict, we will apply the linear regression and logistic regression and evaluate the result based on the training, testing and validation set of the data. Investment and sales prediction using machine learning is an important area of research and application in the finance industry. The ability to accurately predict investment trends and sales patterns can help businesses make informed decisions about their future investments and sales strategies.

ACKNOWLEDGMENT

The satisfaction and euphoria that successful completion of any project is incomplete without the mention of people who made it possible, whose constant guidance and encouragement made my effort fruitful.

First and foremost, I ought to pay our due regards to this institute, which provided me a platform and gave an opportunity to display my skills through the medium of Technical Seminar. I express our heartfelt thanks to our beloved principal **Dr. B S Nagabhushana**, **Don Bosco Institute of Technology**, Bangalore for his encouragement and providing me with the infrastructure.

I express my deep sense of gratitude and thanks to **Dr. K B ShivaKumar, Head of the Department, Computer Science and Engineering** for extending his valuable insight and suggestions offered during this Technical Seminar.

I express my acknowledgement to my seminar coordinator and guide **Prof. Komala D, Asst.Prof., Dept of CSE,** for extending their direction, support, guidance, and assistance which consequently resulted in getting the seminar work completed successfully.

Last but not the least I would like to thank teaching and non-teaching staff for their cooperation extended during the completion of the Technical Seminar.

PREKSHA M [1DB19CS108]

CONTENTS

Chapter 1

COMPANY PROFILE

1.1 INTRODUCTION

Inventeron Technologies and Business Solution LLP is an Indian based engineering and Electronics Company headquartered in Bangalore, Karnataka, India. It is both product and service-oriented software company having its products in wireless communication Technology and provides quality service toits valuable clients in its domain.

1.2 HISTORY

- The company was legally registered in the year 2013, but it made its humble beginning in the year 2012 with a team of six members. In the beginning the team started designing some protocols for wireless communication with a range up to 4 to 5 km line of sight. The company handled various projects and successfully completed them satisfying the client requirement. After the successful completion of the project and achieving customer satisfaction the number of clients increased who sincerely served with respect and faith
- That is how the company started generating the revenue. Even though the team members were experts in embedded electronics, Java, Dot net and android, the company simultaneously established itself to develop websites and few latest apps based on the client requirement. The company was later registered on 24th December 2013 and established a well-equipped office space with good R&D unit and supporting infrastructure. It started recruiting people with great skills and expertise for different domains of company and started working with new hopes and enthusiasm. Presently the company have around 25 employees including all the departments like Embedded, Java, Dot net, android, Testing, PCB design, IOT and so on. The company is working with many Industrial projects in different domains and working for its own product
- AiRobosoft is located at Hebbal, Bengaluru, Karnataka. They are a community of Data Scientist, Robotics & Electronics Engineers, experts in Machine Learning and more, collaborated together to work on fascinating futuristic technologies ensuring safety and ethics empowering humans to overcome critical challenges Artificial intelligence could be one of humanity's most useful inventions. They research and build safe AI systems that learn how to solve problems and advance scientific discovery for all.

- There vision is to develop highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. They will attempt to directly build safe and beneficial AI, but will also consider there mission fulfilled if there work aids others to achieve this outcome.

1.3 COMPANY STRATEGY

- **Purpose:** To be a leader in the software Industry by providing enhanced services, relationship and profitability.
- **Vision:** To provide quality services that exceeds the expectations of our esteemed customers.
- **Mission:** To build long term relationships with our customers and clients and provide exceptional customer services by pursuing business through innovation and advanced technology.
- **Core values:**
- To incorporate good business practices in order to achieve customer satisfaction and treating the customers with respect and faith.
- To grow through creativity, invention and innovation.
- To integrate honesty, integrity and business ethics into all aspects of the business functioning.
- **Goals:**
- To improve, grow and become more efficient in the field electronics engineering and software development and develop a strong base of key clients.
- To understand customer requirements and fulfill them.
- Increase the assets and investments of the organization to support the development of services and expansion of the organization.
- To increase the productivity and improve the customer service satisfaction.
- PCB design, IOT and so on. The company is working with many Industrial projects in different domains and working for its own products

1.4 COMPANY SERVICES

AIROBOSOFT Product and Services LLP have its own services such as,

- Embedded Applications development
- Web design and development

- IT Service
- Server Maintenance
- Project Management
- Company Products
- AIROBOSOFT Product And Services LLP have it company products like
- Smart Surveillance system
- Safety and Security Systems
- Industrial Automation
- Hone Automation
- Biometrics
- Smart Traffic Systems
- Vehicle Tracking Systems
- Tower Management System
- Education Management Systems
- Hotel Management System
- Personal safety Equipment's
- Wireless Communication Devices
- LED Products
- Water Controlling Units
- Water level controller
- smart ration management system

1.5 DOMAINS

AIROBOSOFT Product and Services LLP have working with several domains like-

- Government
- Food and Beverages
- Health Care
- Outsourcing
- HR Management
- Smart Surveillance system
- Safety and Security Systems
- Industrial Automation

1.6 DEPARTMENTS

Production: Production is the functional area responsible for turning inputs into finished outputs through a series of production processes. The Production Manager is responsible for making sure that the materials required are available at the time of developing the product. The Production manager must make sure the work is carried out smoothly and must supervise procedures for making work more efficient. A product is anything that can be offered to a market that might satisfy a want or need.

Marketing: These are the main section of the market departments:

- Sales department is responsible for the sales and distribution of the products to the different regions.
- Research & Department is responsible for market research and testing new products to make sure that they are suitable to be sold.
- Promotion department decides on the type of promotion method for the products, arranges advertisements and the advertising media used.
- Distribution department distributes the products across the industries.
- Embedded System and Internet of Things (IOT) department.
- Machine learning and web development department.

Chapter 2

INTRODUCTION

Predicting Investment and sales of a company needs time series data of that company and based on that data the model can predict the future sales of that company or product. So, in this research project we will analyse the time series sales data of a company and will predict the sales of the company for the coming quarter and for a specific product. For this kind of project of sales predict, we will apply the linear regression and logistic regression and evaluate the result based on the training, testing and validation set of the data.

A sales analysis report shows the trends that occur in a company's sales volume over time. In its most basic form, a sales analysis report shows whether sales are increasing or declining. At any time during the fiscal year, sales managers may analyse the trends in the report to determine the best course of action. Managers often use sales analysis reports to identify market opportunities and areas where they could increase volume. For instance, a customer may show a history of increased sales during certain periods. This data can be used to ask for additional business during these peak periods. A sales analysis report shows a company's actual sales for a specified period a quarter, a year, or any time frame that managers feel is significant. In larger corporations, sales analysis reports may only contain data for a subsidiary, division or region. A small-business manager may be more interested in breaking sales down by location or product. Some small, specialized businesses with a single location are compact enough to use general sales data. A sales analysis report may compare actual sales to projected sales. Linear regression and logistic regression is the best machine learning models for this kind of problem where we can easily fit a line of high sale and low sale product, quarters and zone for a product. Also, we need huge amount of data for the training of the model which we can collect from the sales data of any product or company of last 1 or 2 years for any live project. However, for this research project, the description of the dataset which we are going to use for this project is provided in the dataset portion of experimental setup section.

2.1 METHODOLOGY

In this research, linear regression and logistic regression model will be trained and tested for our dataset. For this we will download the sample dataset from the given link in dataset section. The raw data is then under goes for feature selection and feature extraction. After that we will apply machine learning regression models for the training dataset to train the model. This train model will be then tested on test dataset and validation dataset for checking the accuracy of the model.

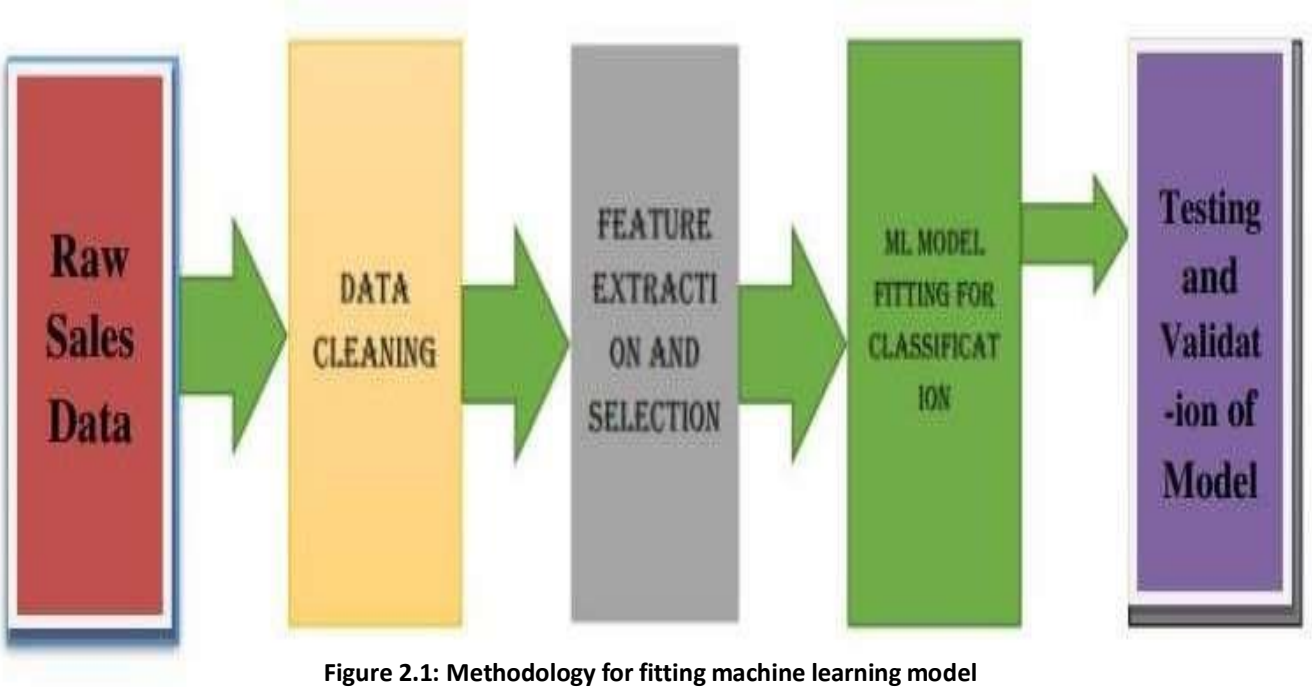

Figure 2.1: Methodology for fitting machine learning model

The method of experiment where we are taking the raw data from our source and will apply some data cleaning methods to make our data smooth.

The trigger model system framework with some classifiers but for this research regression is proposed. Then the most important step is feature extraction and selection will be applied to select best features out of available which are influencing the result more. Then we will apply some machine learning model and compare the results.

2.2 DATASET

We are using the superstore sales data for sales prediction. Sample data that appears in the December Tableau User Group presentation. Note: Geographic locations have been altered to include Canadian locations (provinces / regions).

2.3 EVALUATION MEASURES

Measures such as Classification error, Computational cost, Accuracy can be used for calculating the accuracy of drug discovery using neural network.

2.4 MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. IBM has a rich history with machine learning. One of its own, Arthur Samuel, is credited for coining the term, ‑machine learning‖ with his research (PDF, 481 KB) (link resides outside IBM) around the game of checkers. Robert Nealey, the self-proclaimed checkers master, played the game on an IBM 7094 computer in 1962, and he lost to the computer. Compared to what can be done today, this feat seems trivial, but it’s considered a major milestone in the field of artificial intelligence.

Over the last couple of decades, the technological advances in storage and processing power have enabled some innovative products based on machine learning, such as Netflix’s recommendation engine and self-driving cars. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications o predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase. They will be required to help identify the most relevant business questions and the data to answer them. Machine learning algorithms are typically created using frameworks that accelerate solution development, such as TensorFlow and PyTorch.

Chapter 3

SOFTWARE REQUIREMENTS

Introduction of Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Python is a great general-purpose programming language on its own, but with the help of a few popular libraries (numpy, scipy, matplotlib) it becomes a powerful environment for scientific computing.

Python 2 Published in late 2000, Python 2 signalled a more transparent and inclusive language development process than earlier versions of Python with the implementation of PEP (Python Enhancement Proposal), a technical specification that either provides information to Python community members or describes a new feature of the language.
Additionally, Python 2 included many more programmatic features including a cycle-detecting garbage collector to automate memory management, increased Unicode support to standardize characters, and list comprehensions to create a list based on existing lists. As Python 2 continued to develop, more features were added, including unifying Python's types and classes into one hierarchy in Python version 2.2.

Python 3 Python 3 is regarded as the future of Python and is the version of the language that is currently in development. A major overhaul, Python 3 was released in late 2008 to address and amend intrinsic design flaws of previous versions of the language. The focus of Python 3 development was to clean up the codebase and remove redundancy, making it clear that there was only one way to perform a given task.

Major modifications to Python 3.0 included changing the print statement into a built-in function, improve the way integers are divided, and providing more Unicode support.

At first, Python 3 was slowly adopted due to the language not being backwards compatible with Python 2, requiring people to make a decision as to which version of the language to use. Additionally, many package libraries were only available for Python 2, but as the development team behind Python 3 has reiterated that there is an end of life for Python 2 support, more libraries have been ported to Python 3. The increased adoption of Python 3 can be shown by the number of Python packages that now provide Python 3 support, which at the time of writing includes 339 of the 360 most popular Python packages.

Python 2.7 Following the 2008 release of Python 3.0, Python 2.7 was published on July 3, 2010 and planned as the last of the 2.x releases. The intention behind Python 2.7 was to make it easier for Python 2.x users to port features over to Python 3 by providing some measure of compatibility between the two. This compatibility support included enhanced modules for version 2.7 like unittest to support test automation, argparse for parsing command-line options, and more convenient classes in collections.

Anaconda Python Distribution. Anaconda is an open-source package manager, environment manager, and distribution of the Python and R programming languages. It is commonly used for large-scale data processing, scientific computing, and predictive analytics, serving data scientists, developers, business analysts, and those working in DevOps.

Anaconda offers a collection of over 720 open-source packages, and is available in both free and paid versions. The Anaconda distribution ships with the conda command-line utility. You can learn more about Anaconda and conda by reading the Anaconda Documentation pages.

Why Anaconda?

- User level install of the version of python you want
- Able to install/update packages completely independent of system libraries or admin privileges
- conda tool installs binary packages, rather than requiring compile resources like pip - again, handy if you have limited privileges for installing necessary libraries.

- More or less eliminates the headaches of trying to figure out which version/release of package X is compatible with which version/release of package Y, both of which are required for the install of package Z
- Comes either in full-meal-deal version, with numpy, scipy, PyQt, spyder IDE, etc. or in minimal / alacarte version (miniconda) where you can install what you want, when you need it
- No risk of messing up required system libraries
- Installing on Windows
- Download the Anaconda installer.
- Optional: Verify data integrity with MD5 or SHA-256. More info on hashes
- Double click the installer to launch.
- NOTE: If you encounter any issues during installation, temporarily disable your anti-virus software during install, then re-enable it after the installation concludes. If you have installed for all users, uninstall Anaconda and re-install it for your user only and try again.
- Click Next.
- Read the licensing terms and click I Agree.
- Select an install for ‐Just Me‖ unless you're installing for all users (which requires Windows Administrator privileges).
- Select a destination folder to install Anaconda and click Next.
- NOTE: Install Anaconda to a directory path that does not contain spaces or unicode characters.
- Choose whether to add Anaconda to your PATH environment variable. We recommend not adding Anaconda to the PATH environment variable, since this can interfere with other software. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Command Prompt from the Start Menu.
- Choose whether to register Anaconda as your default Python 3.6. Unless you plan on installing and running multiple versions of Anaconda, or multiple versions of Python, you should accept the default and leave this box checked.
- Click Install. You can click Show Details if you want to see all the packages Anaconda is installing.
- Click Next.
- After a successful installation you will see the ‐Thanks for installing Anaconda‖ .

- You can leave the boxes checked ‐Learn more about Anaconda Cloud‖ and ‐Learn more about Anaconda Support‖ if you wish to read more about this cloud package management service and Anaconda support. Click Finish.
- After your install is complete, verify it by opening Anaconda Navigator, a program that is included with Anaconda. From your Windows Start menu, select the shortcut Anaconda Navigator. If Navigator opens, you have successfully installed Anaconda.
- Installing on macOS
- Download the graphical macOS installer for your version of Python.
- OPTIONAL: Verify data integrity with MD5 or SHA-256. For more information on hashes, see What about cryptographic hash verification?.
- Double-click the .pkg file.
- Answer the prompts on the Introduction, Read Me and License screens.
- On the Destination Select screen, select Install for me only.
- On the Installation Type screen, you may choose to install in another location. The standard install puts Anaconda in your home user directory.
- Click the Install button.
- A successful installation displays the following screen.

What are Jupyter notebook?

The notebook is a web application that allows you to combine explanatory text, math equations, code, and visualizations all in one easily sharable document.

Notebooks have quickly become an essential tool when working with data. You'll find them being used for data cleaning and exploration, visualization, machine learning, and big data analysis. Typically you'd be doing this work in a terminal, either the normal Python shell or with IPython. Your visualizations would be in separate windows, any documentation would be in separate documents, along with various scripts for functions and classes. However, with notebooks, all of these are in one place and easily read together.

How notebook works?

Jupyter notebooks grew out of the IPython project started by Fernando Perez. IPython is an interactive shell, similar to the normal Python shell but with great features like syntax highlighting and code completion. Originally, notebooks worked by sending messages from the web app (the notebook you see in the browser) to an IPython kernel (an IPython application running in the background). The kernel executed the code, then sent it back to the notebook.

The central point is the notebook server. You connect to the server through your browser and the notebook is rendered as a web app. Code you write in the web app is sent through the server to the kernel. The kernel runs the code and sends it back to the server, then any output is rendered back in the browser. When you save the notebook, it is written to the server as a JSON file with a .ipynb file extension.

The great part of this architecture is that the kernel doesn't need to run Python. Since the notebook and the kernel are separate, code in any language can be sent between them. For example, two of the earlier non-Python kernels were for the R and Julia languages. With an R kernel, code written in R will be sent to the R kernel where it is executed, exactly the same as Python code running on a Python kernel. IPython notebooks were renamed because notebooks became language agnostic. The new name Jupyter comes from the combination of Julia, Python, and R. If you're interested, here's a list of available kernels.

Another benefit is that the server can be run anywhere and accessed via the internet. Typically you'll be running the server on your own machine where all your data and notebook files are stored. But, you could also set up a server on a remote machine or cloud instance like Amazon's EC2. Then, you can access the notebooks in your browser from anywhere in the world.

Installing Jupyter Notebook.

By far the easiest way to install Jupyter is with Anaconda. Jupyter notebooks automatically come with the distribution. You'll be able to use notebooks from the default environment. To install Jupyter notebooks in a conda environment, use conda install jupyter notebook. Jupyter notebooks are also available through pip with pip install jupyter notebook.Launching the notebook server.

To start a notebook server, enter jupyter notebook in your terminal or console. This will start the server in the directory you ran the command in. That means any notebook files will be saved in that directory. Typically you'd want to start the server in the directory where your notebooks live. However, you can navigate through your file system to where the notebooks are.

When you run the command (try it yourself!), the server home should open in your browser. By default, the notebook server runs at http://localhost:8888. If you aren't familiar with this, localhost means your computer and 8888 is the port the server is communicating on. As long as the server is still running, you can always come back to it by going to http://localhost:8888 in your browser.

If you start another server, it'll try to use port 8888, but since it is occupied, the new server will run on port 8889. Then, you'd connect to it at http://localhost:8889.

Chapter 4

TASK PERFORMED

4.1 Basics of Python

Python is a high-level, dynamically typed multiparadigm programming language. Python code is often said to be almost like pseudocode, since it allows you to express very powerful ideas in very few lines of code while being very readable. As an example, here is an implementation of the classic quicksort algorithm in Python

Integers:

Integer literals are created by any number without a decimal or complex component.

Floats:

Float literals can be created by adding a decimal component to a number.

Boolean:

Boolean can be defined by typing True/False without quotes

Strings:

String literals can be defined with any of single quotes ('), double quotes (") or triple quotes (''' or """). All give the same result with two important differences**.** If you quote with single quotes, you do not have to escape double quotes and vice-versa. If you quote with triple quotes, your string can span multiple lines.

Complex:

Complex literals can be created by using the notation x + yj where x is the real component and y is the imaginary component.

Variables:

A variable in Python is defined through assignment. There is no concept of declaring a variable outside of that assignment.

Branching (if / elif / else):

Python provides the if statement to allow branching based on conditions. Multiple elif checks can also be performed followed by an optional else clause. The if statement can be used with any evaluation of truthiness.

4.2 List

The first container type that we will look at is the list. A list represents an ordered, mutable collection of objects. You can mix and match any type of object in a list, add to it and remove from it at will.Creating Empty Lists. To create an empty list, you can use empty square brackets or use the list() function with no arguments.

4.3 Loops

In general, statements are executed sequentially: The first statement in a function is executed first, followed by the second, and so on. There may be a situation when you need to execute a block of code several number of times. Programming languages provide various control structures that allow for more complicated execution paths.

For loop

The for loop in Python is used to iterate over a sequence (list, tuple, string) or other iterable objects. Iterating over a sequence is called traversal.

Loop continues until we reach the last item in the sequence. The body of for loop is separated from the rest of the code using indentation.

While loop

The while loop in Python is used to iterate over a block of code as long as the test expression (condition) is true.We generally use this loop when we don't know beforehand, the number of times to iterate.

In while loop, test expression is checked first. The body of the loop is entered only if the test_expression evaluates to True. After one iteration, the test expression is checked again. This process continues until the test_expression evaluates to False. In Python, the body of the while loop is determined through indentation. Body starts with indentation and the first unindented line marks the end. Python interprets any non-zero value as True. None and 0 are interpreted as False.

4.4 String

Strings are used to record the text information such as name. In Python, Strings act as ‑Sequence‖ which means Python tracks every element in the String as a sequence. This is one of the important features of the Python language.

For example, Python understands the string "hello' to be a sequence of letters in a specific order which means the indexing technique to grab particular letters (like first letter or the last letter).

Earlier, while discussing introduction to strings we have introduced the concept of a sequence in Python. In Python, Lists can be considered as the most general version of a "sequence". Unlike strings, they are mutable which means the elements inside a list can be changed!

Lists are constructed with brackets [] and commas separating every element in the list.

4.5 Tuples

The construction of tuples use () with elements separated by commas where in the arguments will be passed within brackets.

When to use Tuples.

You may be wondering, "Why to bother using tuples when they have a few available methods?"

Tuples are not used often as lists in programming but are used when immutability is necessary. While you are passing around an object and if you need to make sure that it does not get changed then tuple become your solution. It provides a convenient source of data integrity.You should now be able to create and use tuples in your programming as well as have a complete understanding of their immutability.

Sets

Sets are an unordered collection of *unique* elements which can be constructed using the set() function

Dictionaries

We have learned about "Sequences" in the previous session. Now, let's switch the gears and learn about "mappings" in Python. These dictionaries are nothing but hash tables in other programming languages.

In this section, we will learn briefly about an introduction to dictionaries and what it consists of:

- Constructing a Dictionary
- Accessing objects from a Dictionary
- Nesting Dictionaries
- Basic Dictionary Methods

Before we dive deep into this concept, let's understand what are Mappings?

Mappings are a collection of objects that are stored by a "key". Unlike a sequence, mapping store objects by their relative position. This is an important distinction since mappings won't retain the order since they have objects defined by a key.

A Python dictionary consists of a key and then an associated value. That value can be almost any Python object.

Introduction to Functions

- Functions will be one of our main building blocks when we construct larger and larger amount of code to solve problems.
- A function groups a set of statements together to run the statements more than once. It allows us to specify parameters that can serve as inputs to the functions.
- Functions allow us to reuse the code instead of writing the code again and again. If you recall strings and lists, remember that len() function is used to find the length of a string. Since checking the length of a sequence is a common task, you would want to write a function that can do this repeatedly at command.
- Function is one of the most basic levels of reusing code in Python, and it will also allow us to start thinking of program design.

4.6 Object Oriented Programming and File I/O

Object Oriented Programming (OOP) is a programming paradigm that allows abstraction through the concept of interacting entities. This programming works contradictory to conventional model and is procedural, in which programs are organized as a sequence of commands or statements to perform.It can be an object as an entity that resides in memory, has a state and it's able to perform some actions.More formally objects are entities that represent instances of a general abstract concept called class**.** In Python, "attributes" are the variables defining an object state and the possible actions are called "methods".In Python, everything is an object also classes and functions.

Creating a class

Suppose we want to create a class, named Person, as a prototype, a sort of template for any number of 'Person' objects (instances).

The following python syntax defines a class:

```
class ClassName(base_classes):
statements
```

Class names should always be uppercase (it's a naming convention)

4.7 Modules and Packages

Modules in Python are simply Python files with the .py extension, which implement a set of functions. Modules are imported from other modules using the import command. Before you go ahead and import modules, check out the full list of built-in modules in the Python Standard library.

When a module is loaded into a running script for the first time, it is initialized by executing the code in the module once. If another module in your code imports the same module again, it will not be loaded twice but once only - so local variables inside the module act as a "singleton" - they are initialized only once.

4.8 Numpy

The NumPy package (read as NUMerical PYthon) provides access to

- a new data structure called arrays which allow
- efficient vector and matrix operations. It also provides
- a number of linear algebra operations (such as solving of systems of linear equations, computation of Eigenvectors and Eigenvalues).

Some background information: There are two other implementations that provide nearly the same functionality as NumPy. These are called –Numeric‖ and –numarray‖:

- Numeric was the first provision of a set of numerical methods (similar to Matlab) for Python. It evolved from a PhD project.
- Numarray is a re-implementation of Numeric with certain improvements (but for our purposes both Numeric and Numarray behave virtually identical).
- Early in 2006 it was decided to merge the best aspects of Numeric and Numarray into the Scientific Python (scipy) package and to provide (a hopefully –final‖) array data type under the module name –NumPy‖.

We will use in the following materials the –NumPy‖ package as provided by (new) SciPy. If for some reason this doesn't work for you, chances are that your SciPy is too old. In that case, you will find that either –Numeric‖ or –numarray‖ is installed and should provide nearly the same capabilities.

4.9 Arrays

We introduce a new data type (provided by NumPy) which is called –array‖. An array *appears* to be very similar to a list but an array can keep only elements of the same type (whereas a list can mix different kinds of objects). This means arrays are more efficient to store (because we don't need to store the type for every element). It also makes arrays the data structure of choice for numerical calculations where we often deal with vectors and matricies.Vectors and matrices (and matrices with more than two indices) are all called –arrays‖ in NumPy.

4.10 Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to work with relational *or* labeled data both. It is a fundamental high-level building block for executing practical, real world data analysis in Python.

pandas is well suited for:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

Key features:

- Easy handling of missing data.
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects.
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the data can be aligned automatically.
- Powerful, flexible group by functionality to perform split-apply-combine operations on data sets.
- Intelligent label-based slicing, fancy indexing, and sub setting of large data sets.
- Intuitive merging and joining data sets.
- Flexible reshaping and pivoting of data sets.
- Hierarchical labeling of axes.
- Robust IO tools for loading data from flat files, Excel files, databases, and HDF5.
- Time series functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging, etc.

4.11 Keras

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.

Up until version 2.3, Keras supported multiple backends, including TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML.As of version 2.4, only TensorFlow is supported. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular,

and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System),and its primary author and maintainer is François Chollet, a Google engineer. Chollet is also the author of the Xception deep neural network model.

Keras contains numerous implementations of commonly used neural-network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code. The code is hosted on GitHub, and community support forums include the GitHub issues page, and a Slack channel.

In addition to standard neural networks, Keras has support for convolutional and recurrent neural networks. It supports other common utility layers like dropout, batch normalization, and pooling.

Keras allows users to productize deep models on smartphones (iOS and Android), on the web, or on the Java Virtual Machine.It also allows use of distributed training of deep-learning models on clusters of Graphics processing units (GPU) and tensor processing units (TPU).

4.12 Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.SciPy makes use of Matplotlib.

Matplotlib was originally written by John D. Hunter. Since then it has had an active development community and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in August 2012 and was further joined by Thomas Caswell.Matplotlib is a Num Focus fiscally sponsored project.

4.13 Handling Missing Data

What is missing data?

Not all missing data is equal. At the heart of the matter, there exists the need to distinguish between two types of missingness:

- **Unknown but existing data**: This is data that we know exists, however, due to sparse or incomplete sampling, we do not actually know the value of it. There is some value there, and it would be useful to try and apply some sort of missing data interpolation technique in order to discover it.

  For example, in 2013 *The New York Times* published a survey of income mobility in the United States. As it happens often in datasets which drill this deep (to a county level), there were several counties for which the newspaper could not trace data. Yet it would be possible, and easy, if it was truly necessary to do so, to interpolate reasonable values for these counties based on data from the surrounding ones, for instance, or based on data from other counties with similar demographic profiles. This is fundamentally speaking, data that *can* be filled by some means.
- **Data that doesn't exist**: data that does not exist at all, in any shape or form.

  For example, it would make no sense to ask the average household income for residents of an industrial park or other such location where no people actually live. It would not *really* make sense to use 0 as a sentinal value in this case, either, because the existance of such a number implies in the first place the existance of people for whom an average can be taken—otherwise in trying to compute an average you are making a divide by zero error! This is, fundamentally speaking, data that *cannot* be filled by any means.
- **Bit patterns**: Embed sentinal values into the array itself. For instance for integer data one might take `0` or `-9999` to signal unknown but existant data. This requires no overhead but can be confusing and oftentimes robs you of values that you might otherwise want to use (like `0` or `-9999`).
- **Masks**: Use a seperate boolean array to "mask" the data whenever missing data needs to be represented. This requires making a second array and knowing when to apply it to the dataset, but is more robust.

Numpy is the linear algebra and vectorized mathematical operation library which underpins the Python scientific programming stack, and its methodologies inform how everything else works. Numpy has masks: these are provided via the `numpy.ma` module. But it has no native bitpatterns! There is still no performant native bitpattern `NA` type available whatsoever.

The lack of a native `NA` type, as is the case in, say, R, is a huge problem for libraries, like Pandas, that should be able to efficiently handle large datasets.

Indeed, Pandas does not use the `numpy.ma` mask**.** Masks are simply non-performant above for the purposes of a library that is expected to be able to handle literally millions of entries entirely

in-memory, as pandas does. Pandas instead defines and uses its own null value sentinels, particularly NaN (np.nan) for null numbers and NaT (a psuedo-native handled under-the-hood); and then allows you to apply your own isnull() mask to your dataset (more on that shortly).

4.14 Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation Y=a+b*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s)

Regression analysis estimates the relationship between two or more variables. Let's understand this with an easy example:

Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using Regression analysis. They are as follows:

- It indicates the significant relationships between dependent variable and independent variable.
- It indicates the strength of impact of multiple independent variables on dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help Market Researchers / Data Analysts / Data Scientists to eliminate and evaluate the best set of variables to be used for building predictive models

What are the features?

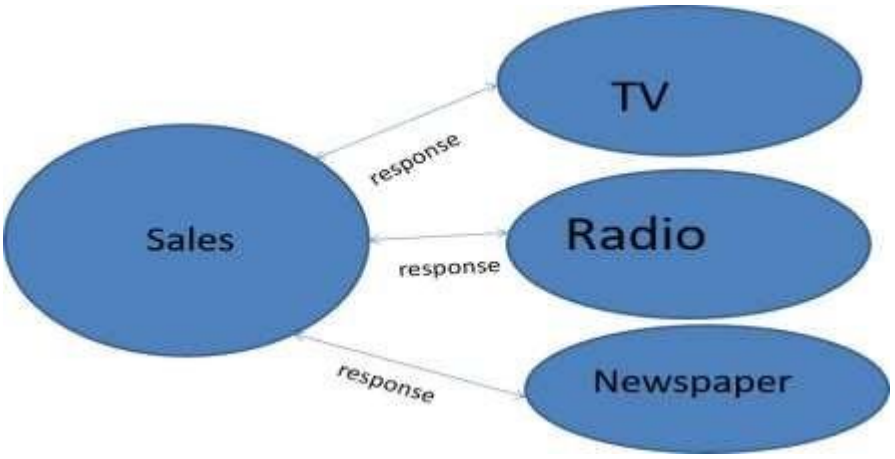

Figure 4.14: Relationship between sales and advertising in media

- TV:Advertising dollars spent on TV for a single product for a single product in a given market.
- Radio: Advertising dollars spent on Radio
- Newspaper: Advertising dollars spent on Newspaper

What is the response

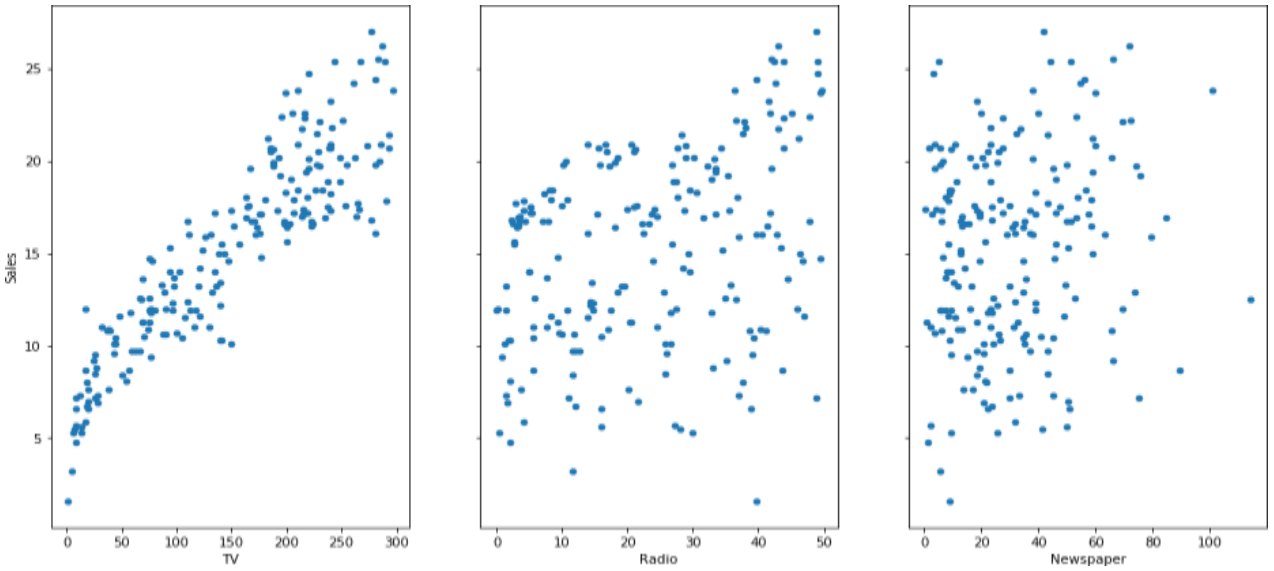

Figure 4.14.1: Relationship between the features and the response using scatterplots

Sales: sales of a single product in a given market (in thousands of widgets)

Questions About the Advertising Data

pretend you work for the company that manufactures and markets this widget. The company might ask you the following: On the basis of this data, how should we spend our advertising money in the future?

These general questions might lead you to more specific questions:

1. Is there a relationship between ads and sales?
2. How strong is that relationship?
3. Which ad types contribute to sales?
4. What is the effect of each ad type of sales?
5. Given ad spending in a particular market, can sales be predicted?

4.15 Simple Linear Regression

Simple Linear regression is an approach for predicting a quantitative response using a single feature (or "predictor" or "input variable"). It takes the following form:
$y=\beta0+\beta1x$y=β0+β1x

What does each term represent?

- yy is the response
- xx is the feature
- $\beta0$β0 is the intercept
- $\beta1$β1 is the coefficient for x

Together, $\beta0$β0 and $\beta1$β1 are called the model coefficients. To create your model, you must "learn" the values of these coefficients. And once we've learned these coefficients, we can use the model to predict Sales!

Estimating ("Learning") Model Coefficients.

Generally speaking, coefficients are estimated using the least squares criterion, which means we find the line (mathematically) which minimizes the sum of squared residuals (or "sum of squared errors")

What elements are present in the diagram?

- The black dots are the observed values of x and y.
- The blue line is the least squares line.
- The red lines are the residuals, which is the distance between the observed values and the least squares line.

How do the model coefficients relate to the least squares line?

- $\beta 0$β0 is the intercept (the value of yy when xx=0)
- $\beta 1$β1 is the slope (the change in yy divided by change in xx)

4.16 The mathematics behind the Least Squares Method.

Take a quick look at the plot created. Now consider each point, and know that each of them have a coordinate in the form (X,Y). Now draw an imaginary line between each point and the current "best-fit" line. We'll call the distanace between each point and the current best-fit line as D. To get a quick image of what we're trying to visualize, take a look at the picture below:

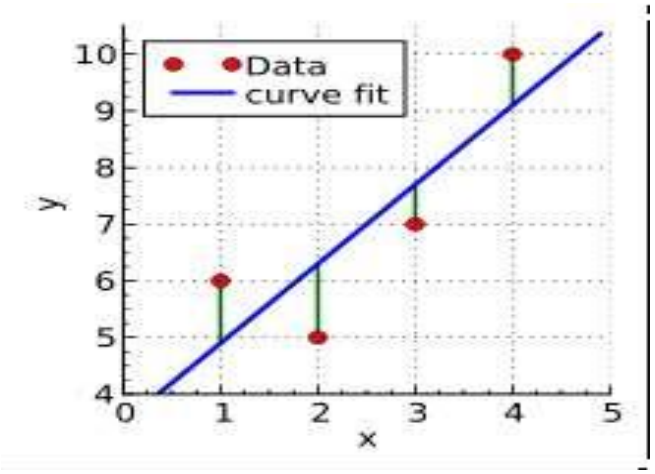

Figure 3.13: Graph of least squares method

Now as before, we're labeling each green line as having a distance D, and each red point as having a coordinate of (X,Y). Then we can define our best fit line as the line having the property were:
$D21$+$D22$+$D23$+$D24$+....+$D2N$D12+D22+D32+D42+ ...+DN2

So how do we find this line? The least-square line approximating the set of points:
(X,Y)1,(X,Y)2,(X,Y)3,(X,Y)4,(X,Y)5,(X,Y)1,(X,Y)2,(X,Y)3,(X,Y)4,(X,Y)5

has the equation:
Y=a0+a1XY=a0+a1X

this is basically just a rewritten form of the standard equation for a line:

$Y=mx+b$Y=mx+b

We can solve for these constants a0 and a1 by simultaneously solving these equations:
$\Sigma Y=a0N+a1\Sigma X$ΣY=a0N+a1ΣX

$\Sigma XY=a0\Sigma X+a1\Sigma X2$ΣXY=a0ΣX+a1ΣX2

These are called the normal equations for the least squares line. There are further steps that can be taken in rearranging these equations to solve for y, but we'll let scikit-learn do the rest of the heavy lifting here.

4.17 Scikit-Learn

Since its release in 2007, scikit-learn has become one of the most popular open source Machine Learning libraries for Python. scikit-learn provides algorithms for Machine Learning tasks including classification, regression, dimensionality reduction, and clustering. It also provides modules for extracting features, processing data, and evaluating models.

Conceived as an extension to the SciPy library, scikit-learn is built on the popular Python libraries NumPy and matplotlib. NumPy extends Python to support efficient operations on large arrays and multidimensional matrices. matplotlib provides visualization tools, and SciPy provides modules for scientific computing.

scikit-learn is popular for academic research because it has a well-documented, easy-to-use, and versatile API. Developers can use scikit-learn to experiment with different algorithms by changing only a few lines of the code. scikit-learn wraps some popular implementations of machine learning algorithms, such as LIBSVM and LIBLINEAR. Other Python libraries, including NLTK, include wrappers for scikit-learn. scikit-learn also includes a variety of datasets, allowing developers to focus on algorithms rather than obtaining and cleaning data.

Licensed under the permissive BSD license, scikit-learn can be used in commercial applications without restrictions. Many of scikit-learn's algorithms are fast and scalable to all but massive datasets. Finally, scikit-learn is noted for its reliability; much of the library is covered by automated tests.

Interpreting Model Coefficients

How do we interpret the TV coefficient (β1β1)?

- A "unit" increase in TV ad spending is associated with a 0.047537 "unit" increase in Sales.
- Or more clearly: An additional $1,000 spent on TV ads is associated with an increase in sales of 47.537 widgets.

Note that if an increase in TV ad spending was associated with a decrease in sales, β1β1 would be negative.

Using the Model for Prediction

Let's say that there was a new market where the TV advertising spend was $50,000. What would we predict for the Sales in that market?

$y=\beta 0+\beta 1x$y=β0+β1x

$y=7.032594+0.047537\times 50$

Plotting the Least Squares Line

make predictions for the smallest and largest observed values of x, and then use the predicted values to plot the least squares line:

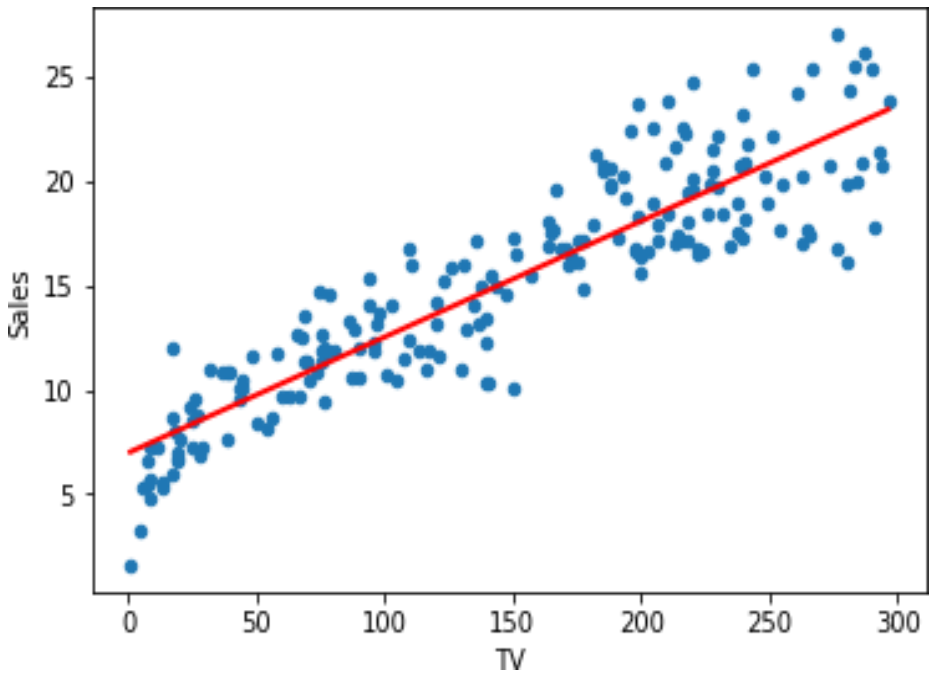

Figure 4.17: Plotting the least squares line

4.18 Multiple Linear Regression

Simple linear regression can easily be extended to include multiple features. This is called multiple linear regression**:**

$y=\beta 0+\beta 1x1+...+\beta nxn$y=β0+β1x1+...+βnxn

Each xx represents a different feature, and each feature has its own coefficient. In this case:

$y=\beta 0+\beta 1\times TV+\beta 2\times Radio+\beta 3\times Newspaper$.

Chapter

RESULTS & CONCLUSION

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.903
Model:	OLS	**Adj. R-squared:**	0.901
Method:	Least Squares	**F-statistic:**	605.4
Date:	Sat, 20 Jul 2019	**Prob (F-statistic):**	8.13e-99
Time:	09:14:19	**Log-Likelihood:**	-383.34
No. Observations:	200	**AIC:**	774.7
Df Residuals:	196	**BIC:**	787.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.6251	0.308	15.041	0.000	4.019	5.232
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012

Omnibus:	16.081	**Durbin-Watson:**	2.251
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27.655
Skew:	-0.431	**Prob(JB):**	9.88e-07
Kurtosis:	4.605	**Cond. No.**	454.

Fig 4.18 Snapshot on Training Dataset

Here is how we interpret the coding:

- **rural** is coded as Area_suburban=0 and Area_urban=0
- **suburban** is coded as Area_suburban=1 and Area_urban=0
- **urban** is coded as Area_suburban=0 and Area_urban=1

	TV	Radio	Newspaper	Sales	Size	IsLarge	Area	Area_suburban	Area_urban
0	230.1	37.8	69.2	22.1	large	1	rural	0	0
1	44.5	39.3	45.1	10.4	small	0	urban	0	1
2	17.2	45.9	69.3	12.0	small	0	rural	0	0
3	151.5	41.3	58.5	16.5	small	0	urban	0	1
4	180.8	10.8	58.4	17.9	large	1	suburban	1	0

Fig 4.19 Snapshot on Sales and Advertising Media on Dataset

Chapter 6

OUTCOMES FROM THE INTERNSHIP

- Explore career alternatives prior to graduation.
- Integrate theory and practice.
- Assess interests and abilities in their field of study.
- Learn to appreciate work and its function in the economy.
- Develop work habits and attitudes necessary for job success.
- Develop communication, interpersonal and other critical skills in the job interview process.
- Build a record of work experience.
- Acquire employment contacts leading directly to a full-time job following graduation from college.
- Identify, write down, and carry out performance objectives related to their job assignment.
- Behaving Professionally.
- Behaving ethically.
- Adapting effectively to changing conditions.
- **Teamwork:** Teamwork is such an important aspect of running successful company and my internships have taught me how to do this on a business level. Teamwork is the ability to work well with other people and be adaptable in order to deal effectively with the demands placed on team, which I have achieved to an expected extent. Employers will ask us to demonstrate this skill in our application by working with other people and cooperating with them to get the best result.
- **Verbal communication**: good verbal communication skills are essential in all areas of business. We need to be able to give clear direction and listen to other workers and customers carefully to ensure that what is done in the workplace is exactly what is needed.
- **Skills:** Aim of the internships teaches us are the skills we need to work in that field. I learned how to take a company's value, needs, and voice. I got opportunity to learn new technologies like IOT and I learnt few languages like embedded c and python. I learnt about my strengths and weakness by creating learning objectives and receiving feedback from our senior's engineers.

Weekly Work:

- We merely took 1 month to develop this website and there are 4 important steps in creation of a website-Planning, Content creation, Website design, Testing and review.
- In the first week, we had introduction about the company, the project we are working on, and some basic requirements were thought to us about the artificial intelligencs and machine learning. After that, the first phase of our project that is planning was done.
- In the second week, we focused on content creation where we collected data related to our project from different sources.
- In the next week, the designing phase was carried out and the front end, back end using different languages and algorithms were developed.
- In the last fourth week, the testing and review of the project was done , the output and accuracy of the developed website was recorded.

REFERENCES

[1] *Akaike, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 6 (1974), 716-723.*

[2] *Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74, 368 (Dec. 1979), 829-836.*

[3] *Diehr, G. and Hoflin, D.R. Approximating the distribution of the sample R2 in best sMarquardt, D.W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 3 (Aug. 1970), 591–612.*

[4] *Mitchell, T. Machine Learning. WCB McGraw-Hill, 1997. 81-82.*

[5] *Ohtani, K. Bootstrapping R2 and adjusted R2 in regression analysis. Economic Modelling, 17, 4 (2000), 473–483.*

[6] *Priddy, K.L. and Keller, Paul E. Artificial Neural Networks: An Introduction. The International Society of Optical Engineering, 2005. 36 - 47.*

[7] *Rencher, A.C. and Pun, F.C. Inflation of R2 in best subset regression. Technometrics, 22, 1 (Feb. 1980), 49–53.*

[8] *Schwarz, G. Estimating the dimension of a model. The Annals of Statistics, 6, 2 (Mar. 1978), 461-464.*

[9] *Studenmund, A.H. Using Econometrics: a Practical Guide. Addison-Wesley, Boston, MA, 2011.*