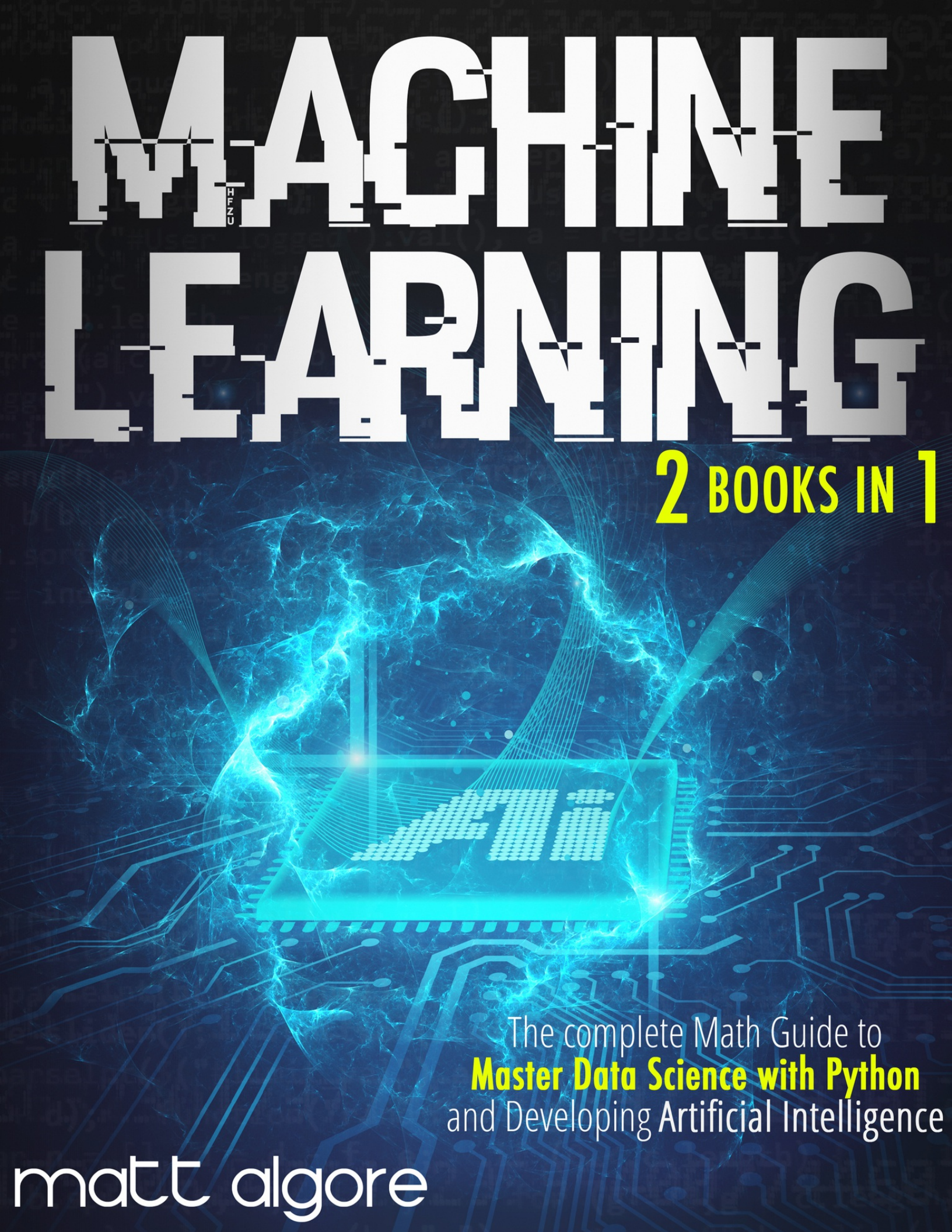
MACHINE LEARNING

2 BOOKS IN 1



The complete Math Guide to
Master Data Science with Python
and Developing Artificial Intelligence

matt algore

MACHINE LEARNING

The complete Math Guide to Master Data Science with Python and Developing Artificial Intelligence

Matt Algore

LEGAL NOTICE

Theory is when you know everything but nothing works. Practice is when everything works but no one knows why. In our lab, theory and practice are combined: nothing works and nobody knows why

(Albert Einstein)

Table of Contents

PART I

Introduction

Machine learning is a computer program that will learn without being explicitly programmed.

Example:

You train a computer to recognise cats and dogs in images. You give examples of images of cats and dogs. You tell the computer that the cats are on the left side of the images and the dogs are on the right side.

Once you've done that, the computer creates some rules that distinguish cats and dogs.

Tests, where the computer can tell the difference between cats and dogs and can't tell the difference in other images, will be run. These tests will show that the computer is learning and its new rules are better than the ones it had at the start.

What Happens When you Try to Teach a Machine to Do Mathematics?

After a few years of training the machine, the only things the computer can do reliably are some elementary things like creating the rules of the game of chess. He best methods that we can follow and focus our time on when it comes to Machine Learning include:

Logistic Regression

Uses a neural network to produce a complex logistic regression system.

Decision Trees

Uses a set of rules to identify a specific tree that classifies the data into a pre-defined set of classes, e.g. a tree that classifies X people into 2 groups, e.g. either black or white.

Naive Bayes

Artificial Neural Networks

Run on a computer, they have a set of nodes and a list of weights for each node which is stored and fed into the system along with the data. The nodes are connected to each other in some way, which is held in the model, and it uses that to see if it falls into a certain category or perspective. This approach often produces more non-linear results than logistic regression systems.

Co-Regression Models

A set of predictive models that can be used to build generalizations of data sets.

Clustering Algorithms

Data can be categorized into groups, e.g. computers can be grouped into departments, each department having a group of students. An algorithm, such as k-means clustering, uses the data to create the groups that result in the best performance.

Why Python and Data Science?

Python is a very powerful and easy-to-use programming language that can be used for many things. It is easy to move on to other things if you get stuck.

Is it Possible to Apply Machine Learning to Aspects of Mathematics?

Yes. Mathematics involves the manipulation of symbols (like numbers and letters). Data sets can be manipulated. Computers can manipulate data.

Can MATLAB in a Computer Do Machine Learning?

No. MATLAB is a very limited programming language, and it is only suitable for very specific things. People use MATLAB because they want to do those things it is good at.

Is it Possible to Use MATLAB in a Computer to Do Machine Learning?

Yes. It is possible to install MATLAB on a computer, install some libraries and modules to use MATLAB to do Machine Learning tasks, or install a C++ compiler so that the MATLAB program can be modified to do Machine Learning tasks.

How Do You Teach a Computer to Do Machine Learning?

There are many ways to teach a computer to do Machine Learning, and the best way will largely depend on the problem that needs to be solved. As a rule, a computer will need lots of examples of something for it to learn from. In Machine Learning, a computer will need examples of the data it is trying to classify and to predict. The best data set is from a data set that is at least as large as the data set that the computer will be predicting from.

Can You Automate Machine Learning?

Yes. Once a computer is taught how to do a Machine Learning task, it is possible to automate that task. Many websites run competitions that run Python scripts that perform Machine Learning tasks. Python on websites is usually written like an entry to a competition. Competition can be run any time there is a problem where a machine is needed to learn a problem and produce an output.

Do Computers Ever Make Mistakes?

Machines are logical and follow the rules they are given. They do not use judgement or language to solve problems. The answers produced are subject to the rules that it has been given.

Who Is to Blame When a Machine Makes a Mistake?

No one can be to blame when a machine makes a mistake. A machine cannot be to blame. A machine can only follow the rules that are given to it.

Relation Between Big Data and Machine Learning (ML)

Big Data is the data too complex and scattered in such a way, that it needs special algorithms and methodologies to process them, so that they can be properly analyzed. The traditional processing system cannot analyze extremely large volumes of data. Machine learning is a refined type of class of algorithms that will process big data. The algorithms like predictive analytics, Text algorithms, Social network mining, etc., play an important role in the Machine Learning process. Machine learning algorithms are based on the algorithms which are capable of analyzing large volumes of data, structured or unstructured. The process to use ML is to prepare the data by cleaning the data to prepare it for the machine learning algorithms, and train the machine learning algorithms to perform the algorithm by making effective combinations of the machine learning algorithms, pre-processing the data to form the input for the machine learning algorithms, passing the output obtained to the Machine Learning algorithms, and training them to perform the function. The ultimate objective is to obtain the best result in the best time, and these are achieved by creating effective algorithms.

Uses of Machine Learning

The applications of Machine Learning include real-time decisions, clinical medicine, fraud detection, search engine results, and oil analysis to name just a few.

Implementing Machine Learning is a very powerful way of making predictions using the algorithms that follow. The algorithms developed have constantly provided better predictions over time so that the performance of the machine learning algorithms is constantly improving.

The predictive power of those algorithms also enables them to be useful in a wide range of areas. The machine learning algorithms will work in all data types including text, images, audio, social media, and financial market data.

Put another way, machine learning techniques are applied to end up with a solution that can reveal meaningful distinctions in data that the naked eye can’t see.

Many different kinds of dependencies (or interrelationships) occur within the data that machine learning algorithms need for their learning. In many examples, this data is made up of large volumes of unstructured textual data. Machine Learning is improved by having all the data that can be used for the learning of the Machine Learning algorithms.

Going forward, Machine learning has been going from strength to strength and it is expected that it will execute tasks that were once only the domain of specialists.

Shared Sensing is an emerging service offered by a mix of established cloud providers and emerging smart city service providers. It refers to the ability for multiple smart city stakeholders to share live information from their networks, resources, and devices for the benefit of the wider community. This will bring numerous benefits, including operational efficiency.

Modern companies recognize the importance of big data for their success, not only to compete with others but also to strengthen their business relation, attract customers. It has already changed the way how the business works, and consumers use the service or buy the products.

For example, Amazon collects data from its customers to provide the most relevant result. It uses the data about what people search for, what they buy,

where they live, etc. Some customers may not want their personal information disclosed as they may look like a potential threat to others. However, most customers accept the terms and conditions and agree to the usage of their data by Amazon.

On the other hand, ISPs use the data from customers to produce revenue. Different ISPs have different ways to produce revenue from customers. For example, your web browser is required to use Google ads. ISPs also use the data from customers to provide different services. For example, some ISPs sell your data to market research companies or insurance companies. These data can be used to verify the email address of customers. This increases the reliability of the email address.

The customers also use data in different ways. But, some people use it in an unethical way by sharing the data with others without the knowledge of others. As the number of people using data grows every day, the methods of sharing the data are improved to make it easier to share the data instantly.

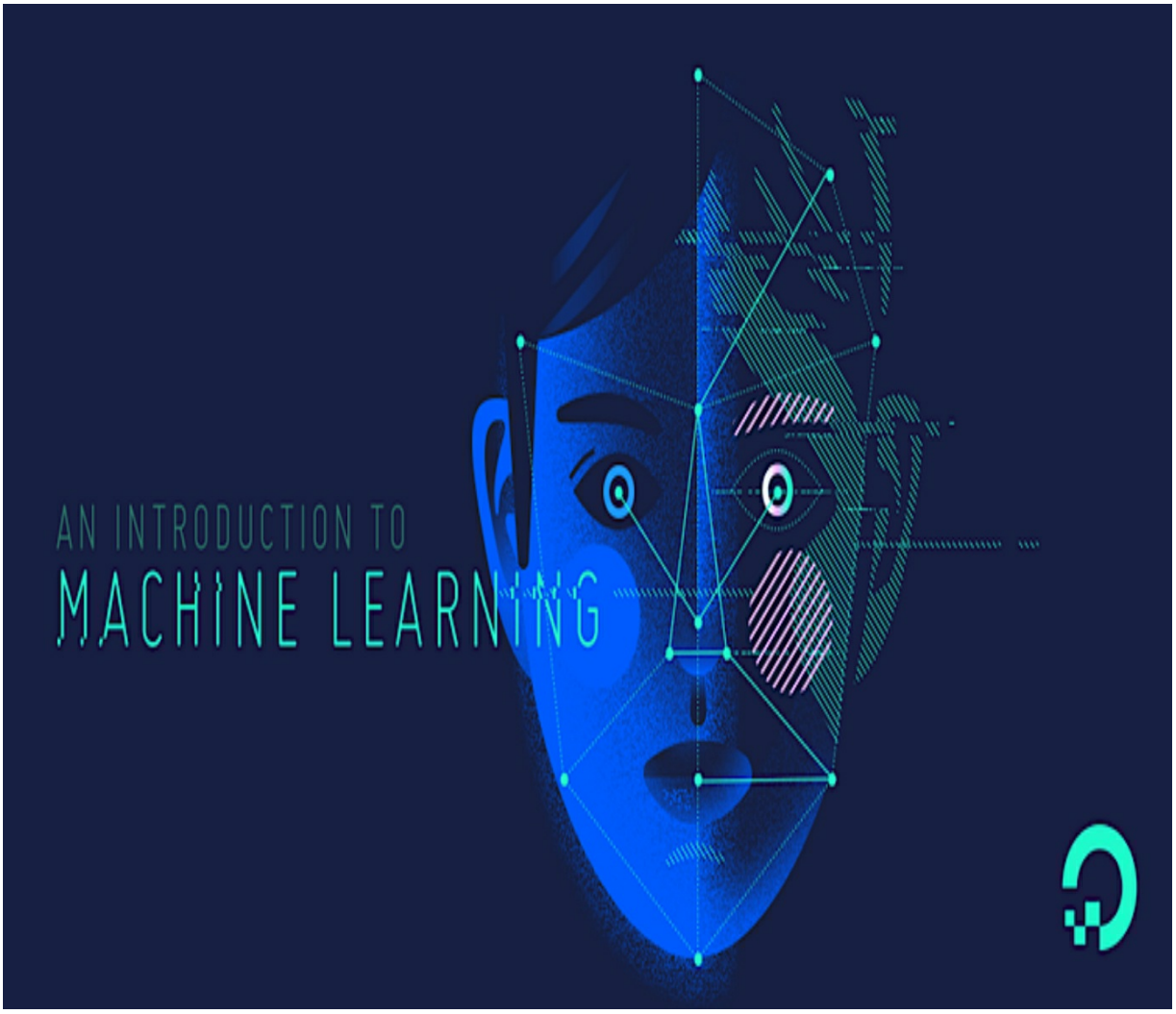
Actual Machine Learning Algorithms

We began the process of Machine Learning by creating an algorithm that had the capability of learning. This created the pattern of breaking the problem into a series of smaller problems to be solved on the data which had been obtained.

There are many techniques that can be applied to Machine Learning, and experts use different techniques for different data sets. The process begins with an algorithm that has the capability of learning to a degree.

This enables the system to learn what works and what does not work. With each tiny step taken, the algorithm can learn if it was possible to align these two groups in a way that the machine learning systems could learn from.

Chapter 1. What Is Machine Learning?



The first thing that we need to take a look at here is the basics of Machine Learning. Machine Learning will be one of the applications of artificial intelligence that can provide a system with the ability to learn, all on its own, without the help of a programmer telling the system what to do. The system can even take this a bit further and improve based on its own experience, and none of this is done with the plan being explicitly programmed in the process. The idea of Machine Learning will be done with a focus on the development

of programs on the computer that can access any data you have, then use that presented data to learn something new, and how you would like it to behave.

There will be a few different applications that we can look at when using Machine Learning. As we start to explore more about what Machine learning can do, you may notice that throughout the years, it has been able to change and develop into something that programmers are going to enjoy working with more than ever. When you want to make your machine or system do a lot of the work independently, without you having to step in and program every step, then Machine Learning is the right option for you.

When it comes to the world of technology, we will find that Machine Learning is pretty unique and can add to a level of fun to the coding that we do. There are already many companies in various industries (which we will talk about in a bit) that will use Machine learning and are already receiving a ton of benefits.

There are many different applications for using Machine Learning, and it is amazing what all we can do with this kind of artificial intelligence. Some of the best methods that we can follow and focus our time on when it comes to Machine Learning include:

1. Research on Statistics

Machine Learning is already making some headway when it comes to the world of IT. You will find that Machine Learning can help you to go through a ton of complex data, looking for the large and essential patterns in the data. Some of the different applications of Machine Learning under this category will include things like spam filtering, credit cards, and search engines.

2. An Analysis of Big Data

Many companies have spent time collecting what is known as Big Data, and now they have to find a way to sort through and learn from that data in a short amount of time. These companies can use this data to learn more about how customers spend money and even help them make important decisions about the future. If we had someone go through and manually do the work, it would take much too long. But with Machine Learning, we can get it all

done. Options like the medical field, election campaigns, and even retail stores have started to turn to Machine Learning to gain some of these benefits.

3. The Financial World

Many financial companies have been able to rely on Machine Learning. Stock trading online, for example, will depend on this kind of work, and we will find that Machine Learning can help with fraud detection, loan approvals, and more.

To help us get going with this one and understand how we can receive the value that we want out of Machine Learning, we have to make sure that we pair the best algorithms with the right processes and tools. If you are using the wrong kind of algorithm to sort through this data, you will get a lot of inaccurate information, and the results will not give you the help you need. Working with the right algorithm, the whole time will make a big difference.

As we are working on some of the models that we want to produce, we will also notice many tools and other processes available for us to work with. We need to make sure that we pick the right one to ensure that the algorithm and the model you are working with will perform the way you would like.

The different tools that are available with Machine learning will include:

1. Comprehensive management and data quality.
2. Automated ensemble evaluation of the model to help see where the best performers will show up.
3. GUIs for helping to build up the models you want and the process flows being built up.
4. Easy deployment of this so that you can quickly get reliable and repeatable results.
5. Interactive exploration of the data and even some visualizations help us view the information easier.
6. A platform that is integrated and end to end to help with the automation of some of the data to decision process that you would like to follow.
7. A tool to compare the different models of Machine learning to help us identify the best one to use quickly and efficiently.

The Benefits of Machine Learning

We also need to take some time to look at a few of the benefits of machine learning. There are many causes why we would want to choose to go with Machine Learning to help our Data Science Project. It is impossible to create some useful algorithms or models that can accurately make predictions out of the data you send through it. There are a lot of other benefits that can come with this as well. Some of the best services that we can see when we decide to work with Machine learning include:

1. Marketing Products Are More Comfortable

When you can reach your customers right where they are looking for you, online and social media, it can increase sales. You can use Machine Learning to figure out what your target audience will respond to, and you can make sure that the products you are releasing work for what the customer wants.

2. Machine Learning Can Help with Accurate Medical Predictions

The medical field is always busy, and it is believed that a lot of the current job openings are going to be left unfilled. Even a regular doctor with no specialties will need to deal with lots of patients throughout the day. Keeping up with all of this can be a hassle. But with the help of Machine learning, we can create a model that can look at images and recognize when something is wrong or not. This can save doctors a lot of time, hassle, and can make them more efficient at their jobs.

This is just one area where Machine learning will be able to help out with the medical field. It can assist with surgeries, take notes for a doctor, look for things in x-rays and other imaging, and even help with front desk operations.

3. Can Make Data Entry Easier

There are times when we need to make sure that all the information is entered into a database efficiently and quickly. If there is a ton of data to sort through and short on time, this can seem like an impossible task. But with Machine

Learning and the tools that come with it, we can get it all done in no time.

4. Helps with Spam Detection

Thanks to some of the learning processes that come with Machine Learning, we find that this can prevent spam. Most of the primary email servers right now will use some form of Machine learning to handle spam and keep it away from your regular inbox.

5. Can Improve the Financial World

Machine Learning can come in and work with many different financial world tasks. It helps with detecting fraud, offering new products to customers, approving loans, and so much more.

6. Can Make Manufacturing More Efficient

Those in the manufacturing world can use Machine Learning to help them be more competent and better at their job. It can figure out when things will be slowing the process down and need to be fixed, and it will look at when a piece of a machine is likely to die out, and so much more.

7. It Requires us with a Better Understanding of the Customer

All companies want to know as much about their customers as possible, ensuring that they can learn how to market to these individuals, what products to offer, and which methods they can take to make the customer as happy as possible.

Supervised Machine Learning

The first type of Machine learning algorithm that we will take a look at is supervised Machine learning. This Machine learning type is the kind where

someone is going to train the system, and the way they do this is by making sure to provide input, with the corresponding output, to the system to know the right answers. You also have to take the time to furnish the feedback into the system, based on whether the system or the machine was accurate in the predictions that it made.

Unsupervised Machine Learning

Now we can move on to the idea of Unsupervised Machine learning and see how this one is going to work compared to supervised learning. With unsupervised Machine learning, we will find that there will be a big difference when compared to the other methods but can train the system how to behave without all of the examples and labeled data along the way.

With unsupervised learning, the model will not be provided with the output for it to be taught how to behave. This is because the goal of this kind of knowledge is that we want the machine to learn what is there, based on the unknown input. The device can know how to do this all on its own, rather than having the programmer come in and do all of the work on it.

Reinforcement Machine Learning

The third type of Machine learning method that we need to take a look at here will be reinforcement Machine learning. This algorithm type is newer than the other two, and it is going to be the one that we work with any time that the presented algorithm has examples, but these examples are not going to have any labels on them at all.

Chapter 2. Giving the Computers the Ability to Learn From Data

We need a programming language to provide instruction to the machine to execute the code to use machine learning. We will learn the basics of the Python language, how to install and launch python. We are also going to learn some Python syntax and some useful tools to run Python. We also cover some necessary Python libraries that useful for machine learning. First of all, why would we use Python and not another programming language?

Why Use Python for Machine Learning?

Python is a programming language extensively used for many reasons. It is a free and open-source language, which means it is accessible to everybody. Although it is free, it is a community-based language, meaning that it is developed and supported by a community that gathers its effort through the internet to improve the language features.

Other reasons people would use Python are:

1. Quality as a readable language with a simple syntax
2. Program portability to any operating system (e.g., Windows, Unix) without or with little modifications
3. Speed of execution: Python does not need compilation and run faster than similar programming languages
4. Component integration which means that Python can be integrated with other programs, can be called from C and C++ libraries, or call another programming language.

Python comes with basic and powerful standard operations and advanced pre-coded libraries like NumPy for numeric programming. Another advantage of Python is automatic memory management and does not require variable and size declaration. Moreover, Python allows developing different applications such as developing Graphical User Interface (GUI), doing numeric programming, do game programming, database programming, internet scripting, and much more.

How to Get Started with Python?

Python is a scripting language, and like any other programming language, needs an interpreter. The latter is a program that executes other language programs. As its name indicates, it works as an interpreter for computer hardware to execute Python programming instructions. Python comes as a software package and can be downloaded from Python's website. When installing Python, the interpreter is usually an executable program. If you use UNIX and LUNIX, Python might be already installed, and it probably is in the /usr directory. Now that you have Python installed let's explore how we can run some necessary code.

To run Python, you can open your operating system's prompt (on Windows,

open a DOS console Window) and type python. If it does not work, you don’t have python in Shell’s path Environment variable. In this case, you should type the full path of the Python executable. On Windows, it should be something similar to C:\Python3.7\python, and in UNIX or LUNIX is installed in the bin folder: /usr/local/bin/python (or /usr/bin/python).

When you launch Python, it provides two lines of information, with the first line is the Python version used as in the example below:

Python 3.7.1 (default, Dec 10 2018, 22:54:23) [MSC v.1915 64 bit (AMD64)]: Anaconda, Inc. on win32

Type "help", "copyright", "credits" or "license" for more information.

>>>

Once a session is launched, Python prompts >>>, which means it is ready. It is prepared to run the line codes you write in. The following is an example of a printing statement:

>>> print ('Hello World!')

Hello World!

>>>

When running Python in an interactive session as we did, it displays the results after >>>, as shown in the example. The code is executed interactively. To exit the interactive Python session, type Ctrl-Z on Windows or Ctrl-D on Unix/Linux machine.

Now we learned how to launch Python and run codes in an interactive session. This is a good way to experiment and test codes. However, the code is never saved, and it needs to be typed again to rerun the statement. To store the code, we need to type it in a file called a module. Files that contain Python statements are called modules. These files have an extension ‘.py.’ The module can be executed simply by typing the module name. A text editor like Notepad++ can be consumed to create the module files. For instance, let's create a module named text.py that prints ‘Hello World’ and calculates 3^2. The file should contain the following statements:

```
print ('Hello World! ')

print ('3^2 equal to ' 3**2)
```

To run this module, in the operating system’s prompt, type the following command line:

```
python test.py
```

If this command line does not work, you should type the full path of Python’s executable and the full path of the test.py file. You can also change the working directory by typing the full cd path of the test.py file, then type python test.py. Changing the working directory to the directory where you saved the modules is a good way to avoid typing the modules' full path every time running the module. The output is:

```
C:\Users>python C:\Users\test.py

Hello World!

3^2 equal to 9
```

When we run the module test.py, the results are displayed in the operating system’s prompt, and they go away as the prompt is closed. To store the results in a file, we can use a shell syntax by typing:

```
python test.py > save.txt
```

The output of test.py is redirected and saved in the save.txt file.

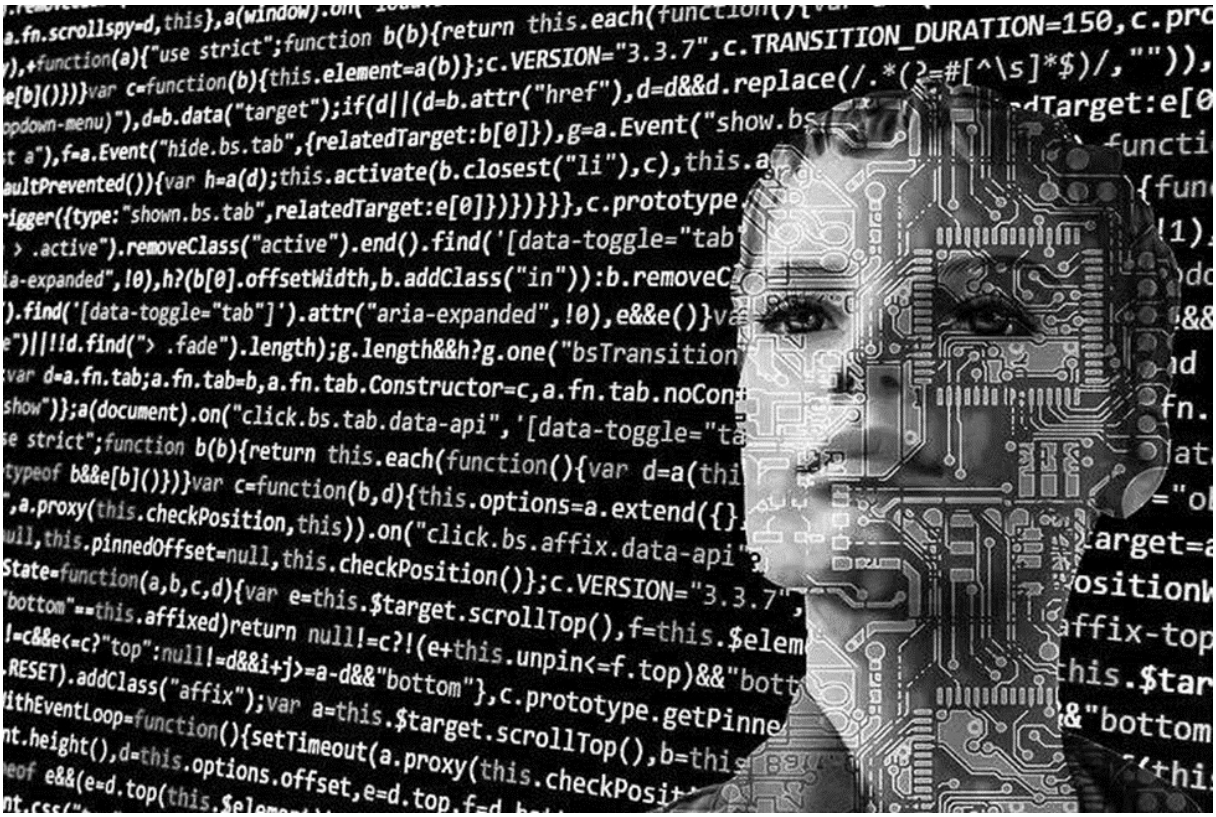
We are going to learn Python syntax. For now, we are going to use the command line to explore Python syntax. We will learn how to set and use some powerful Python programming platforms.

Python Syntax

Before we learn some Python syntax, we will explore the main types of data used in Python and how a program is structured. A plan is a set of modules, which are a series of statements that contain expressions. These expressions create and process objects, which are variables that represent data.

Python Variables

In Python, we can use built-in objects, namely numbers, strings, lists, dictionaries, tuples, and files. Python supports the usual numeric types, the integer, and float, as well as complex numbers. Strings are character chains, whereas lists and dictionaries are ensembles of other objects like a number or a string or other lists or dictionaries. Lists and dictionaries are indexed, and they can be iterated through.



The main difference between lists and dictionaries is how items are stored, and how they can be fetched. Items in a list are ordered and can be fetched by position, whereas they are stored and fetched in dictionaries by key. Tuples like lists are positionally ordered set of objects. Finally, Python also allows creating and reading files as objects. Python provides all the tools and mathematical functions to process these objects.

Python does not require variable declaration, size, or type declaration. Variables are created once they are assigned a value. For example:

```
>>> x=5

>>> print (x) 5

>>> x= 'Hello World! '

Hello World!
```

In the example above, x was assigned a number then it was assigned a string. In fact, Python allows changing the type of variables after they are declared. We can verify the type of any Python object using the type () function.

```
>>> x, y, z=10,'Banana,2.4

>>> print (type(x))

<class 'int '>

>>> print(type(y))

<class 'str '>

>>> print (type(z))

<class 'float '>
```

To declare a string variable, both single and double quotes can be used.

Only alpha-numeric characters and underscores can be used (e.g., A_9). Note that the variable names are case-sensitive and should not start with a number. For instance, price, Price, and PRICE are three different variables. Multiple variables can be affirmed in one line, as seen in the example above.

Chapter 3. Basic Terminology and Notations

Mathematical Notation for Machine Learning

You will realize that mathematical nomenclature and notations go hand in hand throughout your project in your machine learning process. There are various signs, symbols, values, and variables used in mathematics to describe whatever algorithms you may be trying to accomplish.

You will find yourself using some of the mathematical notations within this field of model development. You will find that values that deal with data and the process of learning or memory formation will always take precedence. Therefore, the following six examples are the most commonly used notations. Each of these notations has a description for which its algorithm explains:

Algebra

- To indicate a change or difference: Delta.
- To give the total summation of all values: Summation.
- To describe a nested function: Composite function.
- To indicate Euler's number and Epsilon where necessary.
- To describe the product of all values: Capital pi.

Calculus

- To describe a particular gradient: Nabla.
- To describe the first derivative: Derivative.

- To describe the second derivative: Second derivative.
- To describe a function value as x approaches zero: Limit.

Linear Algebra

- To describe capitalized variables are matrices: Matrix.
- To describe matrix transpose: Transpose.
- To describe a matrix or vector: Brackets.
- To describe a dot product: Dot.
- To describe a Hadamard product: Hadamard.
- To describe a vector: Vector.
- To describe a vector of magnitude 1: Unit vector.

Probability

- The probability of an event: Probability.

Set Theory

- To describe a list of distinct elements: Set.

Statistics

- To describe the median value of variable x: Median.
- To describe the correlation between variables X and Y: Correlation.
- To describe the standard deviation of a sample set: Sample standard deviation.
- To describe the population standard deviation: Standard deviation.

- To describe the variance of a subset of a population: Sample variance.
- To describe the variance of a population value: Population variance.
- To describe the mean of a subset of a population: Sample mean.
- To describe the mean of population values: Population means.

Terminologies Used for Machine Learning

The following terminologies are what you will encounter most often during machine learning. You may be getting into machine learning for professional purposes or even as an artificial intelligence (AI) enthusiast. Anyway, whatever your reasons, the following are categories and subcategories of terminologies that you will need to know and probably understand getting along with your colleagues. Here are machine-learning terms that you need to know:

1. Natural Language Processing (NLP)

Natural language is what you, as a human, use, i.e., human language. By definition, NLP is a way of machine learning where the machine learns your human form of communication. NLP is the standard base for all, if not most, machine languages that allow your device to use human (natural) language. This NLP enables your machine to hear your natural (human) input, understand it, execute it, and then give a data output. The device can realize humans and interact appropriately or as close to appropriate as possible.

There are five primary stages in NLP: machine translation, information retrieval, sentiment analysis, information extraction, and finally, question answering. It begins with the human query, which straight-up leads to machine translation, then through all the four other processes, and finally ending up in question explaining itself. You can now break down these five stages into subcategories as suggested earlier:

Text classification and ranking - This step is a filtering mechanism that determines the class of importance based on relevance algorithms that filter out unwanted stuff such as spam or junk mail. It filters out what needs precedence and the order of execution up to the final task.

Sentiment analysis - This analysis predicts the emotional reaction towards the feedback provided by the machine. Customer relations and satisfaction are factors that may benefit from sentiment analysis.

Document summarization - As the phrase suggests, this is a means of developing short and precise definitions of complex and complicated descriptions. The overall purpose is to make it easy to understand.

Named-Entity Recognition (NER) - This activity involves getting structured and identifiable data from an unstructured set of words. The machine learning process learns to identify the most appropriate keywords, apply those words to the speech context, and try to develop the most appropriate response. Keywords are things like company name, employee name, calendar date, and time.

Speech recognition - An example of this mechanism can easily be appliances such as Alexa. The machine learns to associate the spoken text to the speech originator. The device can identify audio signals from human speech and vocal sources.

It understands Natural language and generation - As opposed to Named-Entity Recognition; these two concepts deal with human to computer and vice versa conversions. Natural language understanding allows the machine to convert and interpret the human form of spoken text into a coherent set of understandable computer format. On the other hand, natural language generation does the reverse function, i.e., transforming the incorrect computer format to the human-understandable audio format.

Machine translation - This action is an automated system of converting one written human language into another human language. Conversion enables people from different ethnic backgrounds and different styles to understand each other. An artificial intelligence entity that has gone through the process of machine learning carries out this job.

2. Dataset

A dataset is a range of variables you can use to test your machine learning's viability and progress. Data is an essential component of your machine

learning progress. It gives results that indicate your development, areas that need adjustments, and tweaking for fine-tuning specific factors. There are three types of datasets:

Training data - As the name suggests, training data is used to predict patterns by letting the model learn via deduction. Due to the enormity of factors to be trained on, yes, there will be more critical factors than others. These features get a training priority. Your machine-learning model will use the more prominent features to predict the most appropriate patterns required. Over time, your model will learn through training.

Validation data - This set is the data used to micro tune the small tiny aspects of the different models at the completion phase. Validation testing is not a training phase; it is a final comparative phase. The data obtained from your validation is used to choose your final model. You get to validate the models' various aspects under comparison and then make a final decision based on this validation data.

Test data - Once you have decided on your final model, test data is a stage that will give you vital information on how the model will handle in real life. The test data will be carried out using an utterly different set of parameters from the ones used during both training and validation. Having the model go through this kind of test data will indicate how your model will handle other types of inputs. You will get answers to questions such as how will the fail-safe mechanism react. Will the fail-safe even come online in the first place?

3. Computer Vision

Computer vision is responsible for the tools, providing a high-level analysis of image and video data. Challenges that you should look out for in computer vision are:

Image classification - This training allows the model to identify and learn what various images and pictorial representations are. The model needs to retain a familiar-looking image to maintain the mind and identify the correct image even with minor alterations such as color changes.

Object detection - Unlike image classification, which detects whether there is

an image in your model field of view, object detection allows it to identify objects. Object identification enables the model to take a large set of data and then frames them to detect pattern recognition. It is akin to facial recognition since it looks for patterns within a given field of view.

Image segmentation - The model will associate a specific image or video pixel with a previously encountered pixel. This association depends on the concept of a most likely scenario based on the frequency of association between a particular pixel and a corresponding specific predetermined set.

Saliency detection - In this case, it will involve that you train and get your model accustomed to increase its visibility. For instance, advertisements are best at locations with higher human traffic. Hence, your model will learn to place itself in positions of maximum social visibility. This computer vision feature will naturally attract human attention and curiosity.

4. Supervised Learning

You achieve supervised learning by teaching the models themselves by using targeted examples. If you wanted to show the models how to recognize a given task, then you would label the dataset for that particular supervised task. You will then present the model with the set of labeled examples and monitor its learning through supervision.

The models get to learn themselves through constant exposure to the correct patterns. You want to promote brand awareness; you could apply supervised learning where the model leans by using the product example and mastering its advertisement art.

5. Unsupervised Learning

This learning style is the opposite of supervised learning. In this case, your models learn through observations. There is no supervision involved, and the datasets are not labeled; hence, there is no correct base value as learned from the supervised method.

Here, through constant observations, your models will get to determine their

right truths. Unsupervised models most often learn through associations between different structures and fundamental characteristics common to the datasets. Since unsupervised learning deals with similar groups of related datasets, they are useful in clustering.

6. Reinforcement Learning

Reinforcement learning teaches your model to strive for the best result always. In addition to only performing its assigned tasks correctly, the model gets rewarded with a treat. This learning technique is a form of encouragement to your model to always deliver the correct action and perform it well or to the best of its ability. After some time, your model will learn to expect a present or favor, and therefore, the model will always strive for the best outcome.

This example is a form of positive reinforcement. It rewards good behavior. However, there is another type of support called negative reinforcement. Negative reinforcement aims to punish or discourage bad behavior. The model gets reprimanded in cases where the supervisor did not meet the expected standards. The model learns that lousy behavior attracts penalties, and it will always strive to do good continually.

7. Neural Networks

The neural network is a concept of interconnected models connected through artificial intelligence. These models though synthetic, have the same level of interactions that is observable between humans. Due to the long period for training and learning models, their interconnectivity level will depend on an automated base system.

Chapter 4. Evaluating Models and Predicting Unseen Data Instances

How Is Python Chosen Over other Tools for Data Science?

Python has been the programming favored language for the regular exercises that information researchers address in a few circumstances and is one of the fundamental information examination systems used around the business. Python is ideal for information researchers who need to actualize numerical

programming into their yield frameworks or join information into electronic applications. It's likewise reasonable for applying frameworks, something that PC researchers additionally need to do.

Suppose the application is written in a natural and average manner. In that case, it is called 'Pythonic.' Past that, Python is consistently prominent for certain limits that have gotten data science planner minds.

Direct Learning

Python's most appealing characteristic is that any individual who has to know it, even amateurs, can do so rapidly and effectively, so that is one reason why disciples favor data science to python. That, in like manner, fits well for dynamic people who contribute brief period mulling over. - most notably, R empowers a truncated learning measure to understand the phonetic structure instead of various lingos.

Data Science Vast Libraries

The actual favored situation of using python for data science would be that python offers associates with a sweeping scope of resources for data mining and PC taking care of. Most data scientists who use Python feel that this universal programming language settles a broad arrangement of troubles by giving inventive approaches to managing late saw as unsolvable issues.

Expandable

Like all various vernaculars, including R, when it relates to flexibility, Python outstands. This is way less unpredictable than the Stata and MATLAB tongues. This backings size gives flexibility and different streets to data scientists to deal with various issues — one explanation YouTube moved to the language. Python can be found across various ventures, enlivening the quick application progression for use examples, taking everything into account.

Colossal Community for Python

One explanation this is so incredibly known to Python is an immediate consequence of her lifestyle. While the data science pack starts to get a handle on it, more people volunteer by building new data science storage facilities. It further stimulates the improvement of the most advanced programming and enlisting strategies open today, which explains many individuals use Python for data science.

The lifestyle is a nearby one, and it has sometimes been more clear to find a reaction to a problematic issue. An actual yield of the web is all you require, so you can quickly find the response to specific issues or talk with someone who may maintain it. On Stack Overflow and Code Mentor, engineers can even associate with their companions.

Why Python and Data Science Mix Well?

Information science incorporates extrapolating helpful information from enormous records, information bases, and information stores. Such outcomes are typically unsorted and difficult to gauge with any sensible precision. ML may connect assorted datasets; however, it needs critical fallacy and force in the calculation. Python addresses that requirement by being a programming language of general use. It enables you to construct CSV performance in a spreadsheet for fast reading of the results. Additionally, more complex outputs of files that can be processed for processing by ML clusters.

For example, climate predictions are based on historical measurements from weather reports over a century old. ML can also render forecasting forecasts more reliable, based on historical weather patterns. Python could do that because code execution is efficient and lightweight, although it is multi-functional. Python can also enable structured, functional programming, and object-oriented patterns, meaning an implementation can be found anywhere.

The Python Package Index now contains over 100,000 libraries, and that number is continuing to grow. As described earlier, Python provides several data science-focused libraries. A simple examination on Google reveals lots of Top 10 Python libraries for lists of data science. The most common library for analyzing data is probably an open-source library named pandas. It is a

highly tuned set of applications that makes Python's data analysis a much-simplified task.

Python has the tool-set to perform a wide range of powerful functions, no matter what experts are looking to be doing with Python, whether prescriptive analytics or predictive causal analytics. It is no wonder that data scientists adopted Python.

Data Science Statistical Learning

Statistical learning is a method for statistical-based interpretation of results, categorized as unsupervised or supervised.

One straightforward approach to explain statistical learning is to evaluate the relationship among predictor variables (features, independent variables) & responses (dependent variable) and to create an objective model that could predict the variable response (Y) based on predictor variables (X).

Inference and Prediction

In cases in which a set of inputs, X is readily accessible. Still, output Y is not understood. We sometimes view f as a black box (not connected to the exact shape of f), as much as it produces precise predictions for Y. It is a foretaste.

There are cases when we want to consider the way Y is influenced when X improves. We want to estimate f in this situation, but our aim is not really to generate forecasts for Y. We're more interested in explaining the connection between X and Y here. But f cannot be viewed as a black box, as we need to know the precise structure. This is inferential. Throughout actual life, you can find various issues going into the environment of assumptions, the environment of inferences, or a mixture of both.

Parametric and Non-Parametric Functions

If we take the statistical model off and attempt to approximate f by measuring the parameters' collection, such methods are considered parametric techniques.

Non-parametric techniques don’t make clear statements regarding the shape of f but rather aim to approximate f that comes as near as possible to the datasets.

Model Interpretability and Prediction Accuracy

Among the various approaches we use to study statistics, others are less versatile or more rigid. If the inference is the target, comfortable and reasonably inflexible mathematical analysis methods have advantages. When we are just involved in modeling, we use accessible modular models.

Model Accuracy Assessing

Estimates do not have a free meal, ensuring that no approach beats all the others for all available data sets. The most widely used factor in the regression framework is the MSE (Mean Squared Error). The most commonly used metric in the classification framework is the uncertainty matrix. The essential property of mathematical learning is that training error may decrease as model variability grows, but the test error does not.

Variance and Bias

Bias is the simplifying premises a designer creates for a smoother understanding of the goal task. Parametric models have a high bias, making them easy to know and more straightforward to understand but less versatile in general. Low-bias ML algorithms are Decision Trees, k-Nearest Neighbors, and Auxiliary Vector Machines. Linear Regression, Conditional Logistic Regression, and Discriminant Analysis are all methods in high-bias ML.

Variance is the rate that the goal role prediction might alter if specific training data were used. There is a large variance of non-parametric equations that provide a lot of variabilities. Logistic and Linear Regression, Linear Discriminant Analysis are techniques for learning machines with small

variances. Decision Trees, k-Nearest Neighbors, and Help Vector Machines are ML algorithms with large variance.

Variance and Bias Relationship

The relation in statistical learning among variance and bias is such that:

- The variance may decline with rising bias.
- Rising variance can reduce bias.

There is an exchange-off between these two considerations and the templates we use, and with our question, the approach we want to customize them seeks various compromises in this trade-off.

Choosing the appropriate degree of versatility in both the classification and regression settings is essential to every predictive learning process's performance. The exchange-off of variance-bias, and the resultant U-shape in the testing mistake, will render this a challenging challenge.

Relation Between Big Data and Machine Learning (ML)

With the amount of data produced by individuals and companies at a skyrocketing speed, several concepts such as big data, deep learning, etc., have arisen. It's very natural to inquire if each other profits from these kinds of stuff. We will explore how big data helps ML to assist in making decisions.

Modern companies recognize the importance of big data, but they also realize that it can be much more efficient when combined with automated processes, and this is precisely where ML's strength falls into the frame. ML systems support businesses in various ways, including maintaining, assessing, and using the data captured far more effectively than ever.

In the general definition, ML is a series of technologies that allow linked computers and machines to know, create, and enhance through various approaches, based on their own experience. All the large companies, major

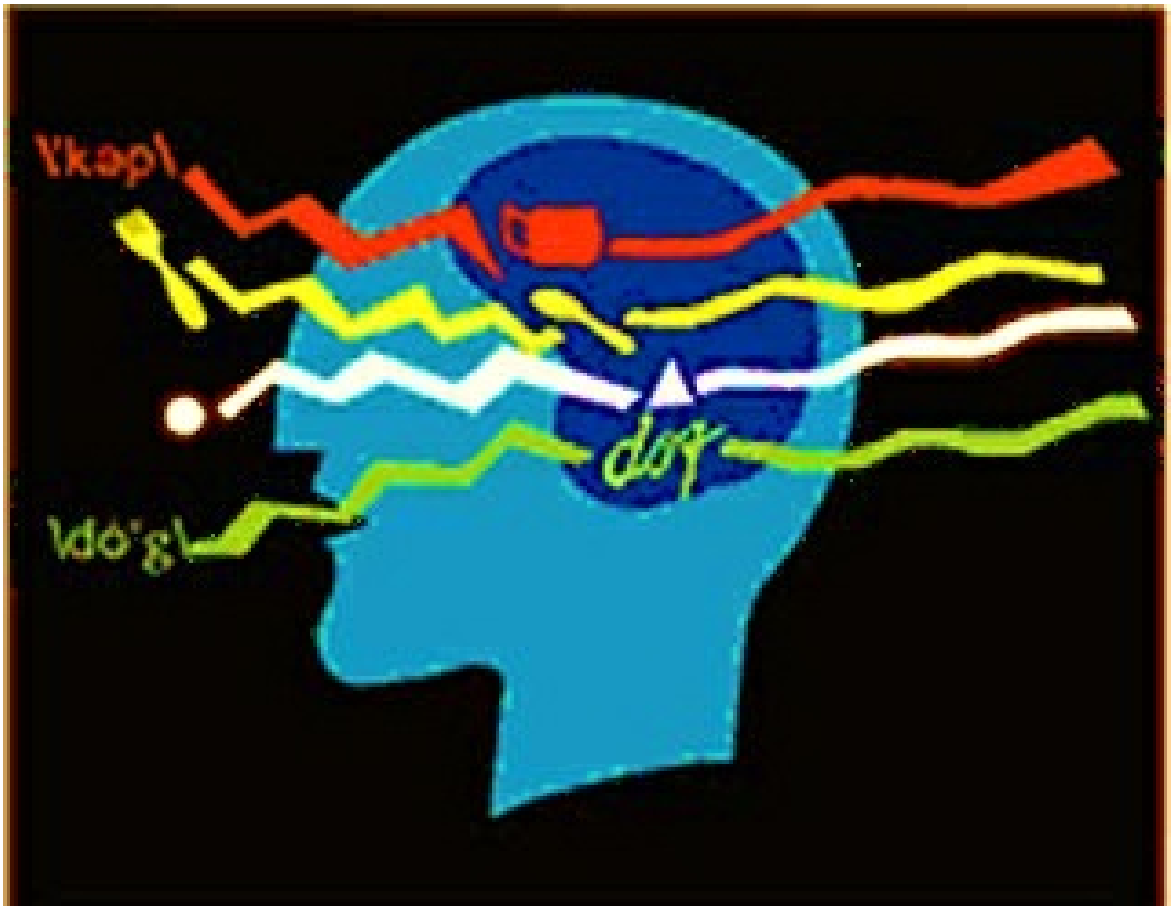
software organizations, and computer scientists are forecasting these days that big data can create a considerable change in the world of machine-learning.

Fundamentally, ML is indeed an advanced type of artificial intelligence designed to learn new information from datasets of its own. It is focused on the premise that machines can learn from results, identify user trends, and make decisions without any human involvement.

Even as ML has been out for decades, models that can analyze more complicated, larger datasets, and generate more reliable data quickly and on a large scale – have become feasible nowadays. By developing such kinds of templates, a company becomes more likely to locate lucrative prospects out.

ML means no previous hypotheses. When ML algorithms are presented with the correct data, they will process the data and recognize trends. You will also use specific findings on other datasets. Such an approach is typically applied to high-dimension datasets. This ensures the more details you will provide, the more reliable the predictions would be. So here's precisely where big data 's influence falls in.

Chapter 5. Building Good Training Datasets



We introduce managing data as a Pandas dataframe and typical exploratory data analysis (EDA) techniques for querying your data.

As a crucial part of data inspection, EDA summarizes your dataset's critical characteristics in preparation for further processing. This includes understanding the data's shape and distribution, scanning for missing values, learning which features are most relevant based on correlation, and familiarizing yourself with the dataset's contents. Gathering this intel helps

inform algorithm selection and highlight aspects of the dataset that require cleaning to prepare for further processing.

Using Pandas, we can use a range of simple techniques to summarize the data and additional options to visualize the data using Seaborn and Matplotlib.

Let’s begin by importing Pandas, Seaborn, and Matplotlib inline using the following code in Jupyter Notebook.

```
import pandas as pd

import seaborn as sns

%matplotlib inline
```

Note that using the inline feature of Matplotlib, we can display plots directly below the applicable code cell within Jupyter Notebook or other frontends.

Import Dataset

Datasets can be imported from various sources, including internal and external files, and random self-generated datasets called blobs.

The following sample dataset is an external dataset downloaded from Kaggle, called the Berlin Airbnb dataset. This data was scraped from Airbnb and contained detailed accommodation listings in Berlin, including location, price, and reviews.

Feature	Data Type	Continuous/Discrete
id	Integer	Discrete
name	String	Discrete
host_id	Integer	Discrete
host_name	String	Discrete
neighbourhood_group	String	Discrete
neighbourhood	String	Discrete
latitude	String	Discrete
longitude	String	Discrete
room_type	String	Discrete
price	Integer	Continuous
minimum_nights	Integer	Continuous
number_of_reviews	Integer	Continuous
last_review	TimeDate	Discrete
reviews_per_month	Floating-point	Continuous
calculated_host_listings_count	Integer	Continuous
availability_365	Integer	Continuous

Overview of the Berlin Airbnb dataset

After registering a free account and logging into Kaggle, download the dataset as a zip file. Then, unzip the downloaded file called listings.csv and import it into Jupyter Notebook as a Pandas dataframe using pd.read_csv.

```
df = pd.read_csv('~/Downloads/listings.csv')
```

Note that you'll need to assign a variable name to store the dataset for ongoing reference. Common variable names for dataframes are "df" or "dataframe," but you can also choose another variable name on the condition that it fits with Python's naming conventions

Remember that your dataset's file path will vary depending on its saved location and your computer's operating system. If saved to Desktop on Windows, you would import the .csv file using a structure similar to this example:

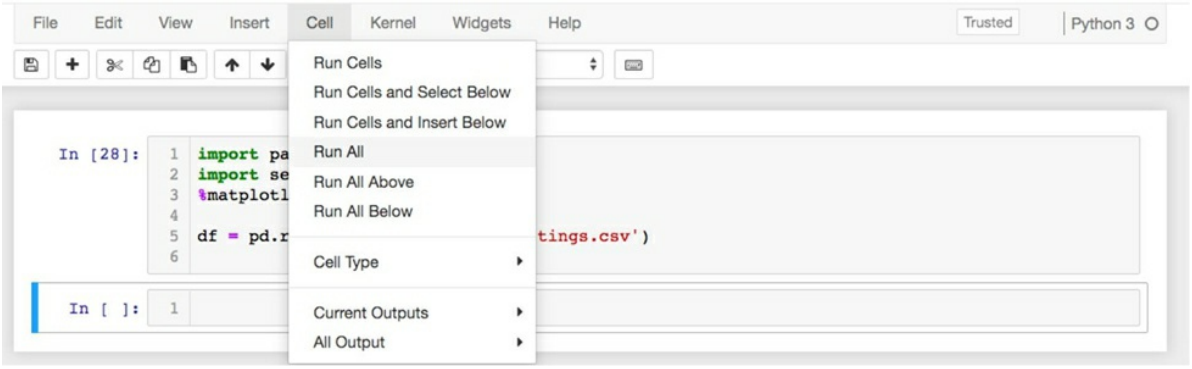
```
df = pd.read_csv('C:\\Users\\John\\Desktop\\listings.csv')
```

Preview the Dataframe

We can now use the Pandas' head() command to preview the dataframe in Jupyter Notebook. The head() command must come after the variable name of the dataframe, which is df.

df.head()

To preview the dataframe, run the code by using right-click, and selecting "Run" or navigating from the Jupyter Notebook menu: Cell > Run All



Run All from the navigation menu

After the code has run, Pandas will populate the imported dataset as a dataframe, as shown in the screenshot.

```
import pandas as pd
import seaborn as sns
%matplotlib inline

df = pd.read_csv('~/Downloads/listings.csv')

df.head()

```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price
0	2015	Berlin-Mitte Value! Quiet courtyard/very central	2217	Ian	Mitte	Brunnenstr. Süd	52.534537	13.402557	Entire home/apt	60
1	2695	Prenzlauer Berg close to Mauerpark	2986	Michael	Pankow	Prenzlauer Berg Nordwest	52.548513	13.404553	Private room	17
2	3176	Fabulous Flat in great Location	3718	Britta	Pankow	Prenzlauer Berg Südwest	52.534996	13.417579	Entire home/apt	90
3	3309	BerlinSpot Schöneberg near KaDeWe	4108	Jana	Tempelhof - Schöneberg	Schöneberg-Nord	52.498855	13.349065	Private room	26
4	7071	BrightRoom with sunny greenview!	17391	Bright	Pankow	Helmholtzplatz	52.543157	13.415091	Private room	42

Previewing a dataframe in Jupyter Notebook using head()

Notice that the first row (ID 2015, located in Mitte) is indexed at position 0 of the dataframe. The fifth row, meanwhile, is indexed at position 4. The indexing of Python elements starts at 0, which means you will need to subtract one from the actual number of rows when calling a specific row from the dataframe.

The dataframe's columns, while not labeled numerically, abide by this same logic. The first column (ID) is indexed at 0, and the fifth column (neighbourhood_group) is indexed at 4. This is a fixed feature of working in Python and something to keep in mind when calling specific rows or columns.

By default, head() displays the first five rows of the dataframe, but you can expand the number of rows by specifying n number of rows inside parentheses, as demonstrated in Figure 9.

```
df = pd.read_csv('~/Downloads/listings.csv')

df.head(10)
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights
0	2015	Berlin-Mitte Value! Quiet courtyard/very central	2217	Ian	Mitte	Brunnenstr. Süd	52.534537	13.402557	Entire home/apt	60	4
1	2695	Prenzlauer Berg close to Mauerpark	2986	Michael	Pankow	Prenzlauer Berg Nordwest	52.548513	13.404553	Private room	17	2
2	3176	Fabulous Flat in great Location	3718	Britta	Pankow	Prenzlauer Berg Südwest	52.534996	13.417579	Entire home/apt	90	62
3	3309	BerlinSpot Schöneberg near KaDeWe	4108	Jana	Tempelhof - Schöneberg	Schöneberg-Nord	52.498855	13.349065	Private room	26	5
4	7071	BrightRoom with sunny greenview!	17391	Bright	Pankow	Helmholtzplatz	52.543157	13.415091	Private room	42	2
5	9991	Geourgeous flat - outstanding views	33852	Philipp	Pankow	Prenzlauer Berg Südwest	52.533031	13.416047	Entire home/apt	180	6
6	14325	Apartment in Prenzlauer Berg	55531	Chris + Oliver	Pankow	Prenzlauer Berg Nordwest	52.547846	13.405562	Entire home/apt	70	90
7	16401	APARTMENT TO RENT	59666	Melanie	Friedrichshain-Kreuzberg	Frankfurter Allee Süd FK	52.510514	13.457850	Private room	120	30
8	16644	In the Heart of Berlin - Kreuzberg	64696	Rene	Friedrichshain-Kreuzberg	nördliche Luisenstadt	52.504792	13.435102	Entire home/apt	90	60
9	17409	Downtown Above The Roofs In	67590	Wolfram	Pankow	Prenzlauer Berg Südwest	52.529071	13.412843	Private room	45	3

Previewing the first ten rows of a dataframe

The argument head(10) is used to preview the first ten rows of the dataframe. You can also view columns concealed to the right by scrolling to the right inside Jupyter Notebook. Regarding rows, you can only preview what's specified in the code.

Lastly, you will sometimes see n= inserted inside the head(), an alternative method to specify n number of previewed rows.

Example Code:

df.head(n=10)

Dataframe Tail

The inverse operation of previewing the top n rows of the dataframe is the tail() method, displaying the bottom n rows of the dataframe. Below, we can

see an example of previewing the dataframe using tail(), which by default also displays five rows. Again, you will need to run the code to view the output.

```
1 import pandas as pd
2 import seaborn as sns
3 %matplotlib inline
4
5 df = pd.read_csv('~/Downloads/listings.csv')
6
7 df.tail()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights
22547	29856708	Cozy Apartment right in the center of Berlin	87555909	Ulisses	Mitte	Brunnenstr. Süd	52.533865	13.400731	Entire home/apt	60	2
22548	29857108	Altbau/ Schöneberger Kiez / Schlafsofa	67537363	Jörg	Tempelhof - Schöneberg	Schöneberg-Nord	52.496211	13.341738	Shared room	20	1
22549	29864272	Artists loft with garden in the center of Berlin	3146923	Martin	Pankow	Prenzlauer Berg Südwest	52.531800	13.411999	Entire home/apt	85	3
22550	29866805	Room for two with private shower / WC	36961901	Arte Luise	Mitte	Alexanderplatz	52.520802	13.378688	Private room	99	1
22551	29867352	Sunny, modern and cozy flat in Berlin Neukölln :)	177464875	Sebastian	Neukölln	Schillerpromenade	52.473762	13.424447	Private room	45	5

Previewing the last five rows of a dataframe using tail()

Find Row Item

While the head and tail commands are useful for gaining a general idea of the dataframe's basic structure, these methods aren't practical for finding an individual or multiple rows in the middle of a large dataset.

To retrieve a specific row or a sequence of rows from the dataframe, we can use the iloc[] command as demonstrated.

df.iloc[99]

```
1 import pandas as pd
2 import seaborn as sns
3 %matplotlib inline
4
5 df = pd.read_csv('~/Downloads/listings.csv')
6
7 df.iloc[99]
```

```
id                                                151249
name                          Quiet, Terrasse, cats, baby equiped
host_id                                           728298
host_name                                          Julia
neighbourhood_group                  Friedrichshain-Kreuzberg
neighbourhood                         Südliche Friedrichstadt
latitude                                        52.4968
longitude                                       13.4168
room_type                                 Entire home/apt
price                                             81
minimum_nights                                     1
number_of_reviews                                 10
last_review                                 2018-09-16
reviews_per_month                                0.26
calculated_host_listings_count                     1
availability_365                                 342
Name: 99, dtype: object
```

Finding a row using .loc[ ]

Here, df.iloc[99] is used to retrieve the row indexed at position 99 in the dataframe, ID 151249 (a listing located in the neighborhood group Friedrichshain-Kreuzberg).

Shape

A quick method to inspect the size of the dataframe is the shape command, which yields rows and columns in the dataframe. This is useful because the dataset's size is likely to change as you remove missing values, recreate features, or delete features.

To doubt the number of rows and columns in the dataframe, you can use the shape command preceded by the dataset's name (parenthesis are not used with this command).

df.shape

```
1  import pandas as pd
2  import seaborn as sns
3  %matplotlib inline
4
5  df = pd.read_csv('~/Downloads/listings.csv')
6
7  df.shape
```

```
(22552, 16)
```

Inspecting the shape (number of rows and columns) of the dataframe

In the case of this dataframe, there are 22,552 rows and 16 columns.

Columns

Another useful command is columns, which prints the dataframe's column titles. This is useful for copying and pasting columns back into the code or clarifying the name of specific variables.

df.columns.

```
1  import pandas as pd
2  import seaborn as sns
3  %matplotlib inline
4
5  df = pd.read_csv('~/Downloads/listings.csv')
6
7  df.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

Print columns

Describe

The describe() method is convenient for generating a summary of the dataframe's mean, standard deviation, and IQR (interquartile range) values. This method performs optimally with continuous values (integers or floating-point numbers that can be aggregated).

df.describe()

```
import pandas as pd
import seaborn as sns
%matplotlib inline

df = pd.read_csv('~/Downloads/listings.csv')

df.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings
count	2.255200e+04	2.255200e+04	22552.000000	22552.000000	22552.000000	22552.000000	22552.000000	18638.000000	22552
mean	1.571560e+07	5.403355e+07	52.509824	13.406107	67.143668	7.157059	17.840679	1.135525	1
std	8.552069e+06	5.816290e+07	0.030825	0.057964	220.266210	40.665073	36.769624	1.507082	3
min	2.015000e+03	2.217000e+03	52.345803	13.103557	0.000000	1.000000	0.000000	0.010000	1
25%	8.065954e+06	9.240002e+06	52.489065	13.375411	30.000000	2.000000	1.000000	0.180000	1
50%	1.686638e+07	3.126711e+07	52.509079	13.416779	45.000000	2.000000	5.000000	0.540000	1
75%	2.258393e+07	8.067518e+07	52.532669	13.439259	70.000000	4.000000	16.000000	1.500000	1
max	2.986735e+07	2.245081e+08	52.651670	13.757642	9000.000000	5000.000000	498.000000	36.670000	45

Using the describe method to summarize the dataframe

By default, describe() excludes columns containing non-numeric values and instead provides a statistical summary of those columns containing numeric values. However, it's also possible to run this command on non-numerical values by adding the argument include='all' within parentheses to obtain the summary statistics of both numeric and non-numeric columns (where applicable).

df.describe(include='all').

```
import pandas as pd
import seaborn as sns
%matplotlib inline

df = pd.read_csv('~/Downloads/listings.csv')

df.describe(include='all')
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minim
count	2.255200e+04	22493	2.255200e+04	22526	22552	22552	22552.000000	22552.000000	22552	22552.000000	225
unique	NaN	21873	NaN	5997	12	136	NaN	NaN	3	NaN	
top	NaN	Berlin Wohnung	NaN	Anna	Friedrichshain-Kreuzberg	Tempelhofer Vorstadt	NaN	NaN	Private room	NaN	
freq	NaN	14	NaN	216	5497	1325	NaN	NaN	11534	NaN	
mean	1.571560e+07	NaN	5.403355e+07	NaN	NaN	NaN	52.509824	13.406107	NaN	67.143668	
std	8.552069e+06	NaN	5.816290e+07	NaN	NaN	NaN	0.030825	0.057964	NaN	220.266210	
min	2.015000e+03	NaN	2.217000e+03	NaN	NaN	NaN	52.345803	13.103557	NaN	0.000000	
25%	8.065954e+06	NaN	9.240002e+06	NaN	NaN	NaN	52.489065	13.375411	NaN	30.000000	
50%	1.686638e+07	NaN	3.126711e+07	NaN	NaN	NaN	52.509079	13.416779	NaN	45.000000	
75%	2.258393e+07	NaN	8.067518e+07	NaN	NaN	NaN	52.532669	13.439259	NaN	70.000000	
max	2.986735e+07	NaN	2.245081e+08	NaN	NaN	NaN	52.651670	13.757642	NaN	9000.000000	50

All variables added to the description

Having consolidated methods to inspect and query the size of the dataframe using Pandas, we'll now move on to generating visual summaries of the data using Seaborn and Matplotlib.

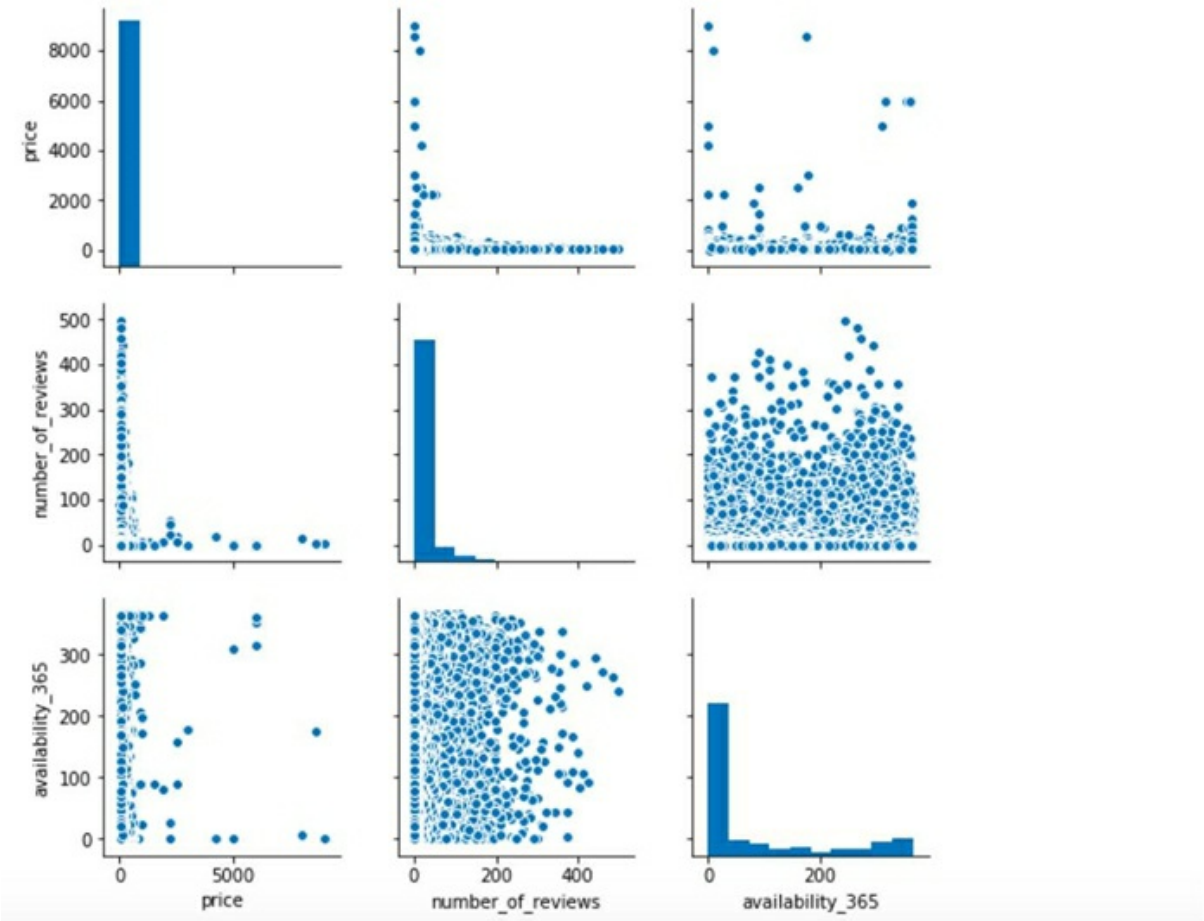
Pairplots

One of the most popular exploratory techniques for understanding patterns between two variables is the pairplot. A pairplot takes the form of a 2-D or 3-D grid of plots that plot variables against other variables taken from the dataframe, as shown in Figure 16.

sns.pairplot(df,vars=['price','number_of_reviews','availability_365'])

```
1  import pandas as pd
2  import seaborn as sns
3  %matplotlib inline
4
5  df = pd.read_csv('~/Downloads/listings.csv')
6  sns.pairplot(df,vars=['price','number_of_reviews','availability_365'])
7
```

```
<seaborn.axisgrid.PairGrid at 0x1a1a684240>
```



Example of a pairplot grid based on three chosen variables

Using a pairplot from Seaborn, we've plotted three chosen variables against each other, which helps us to understand relationships and variance between those variables. When plotted against other variables (multivariant), the visualization takes the form of a scatterplot, and when plotted against the same variable (univariant), a simple histogram is generated.

Heatmaps

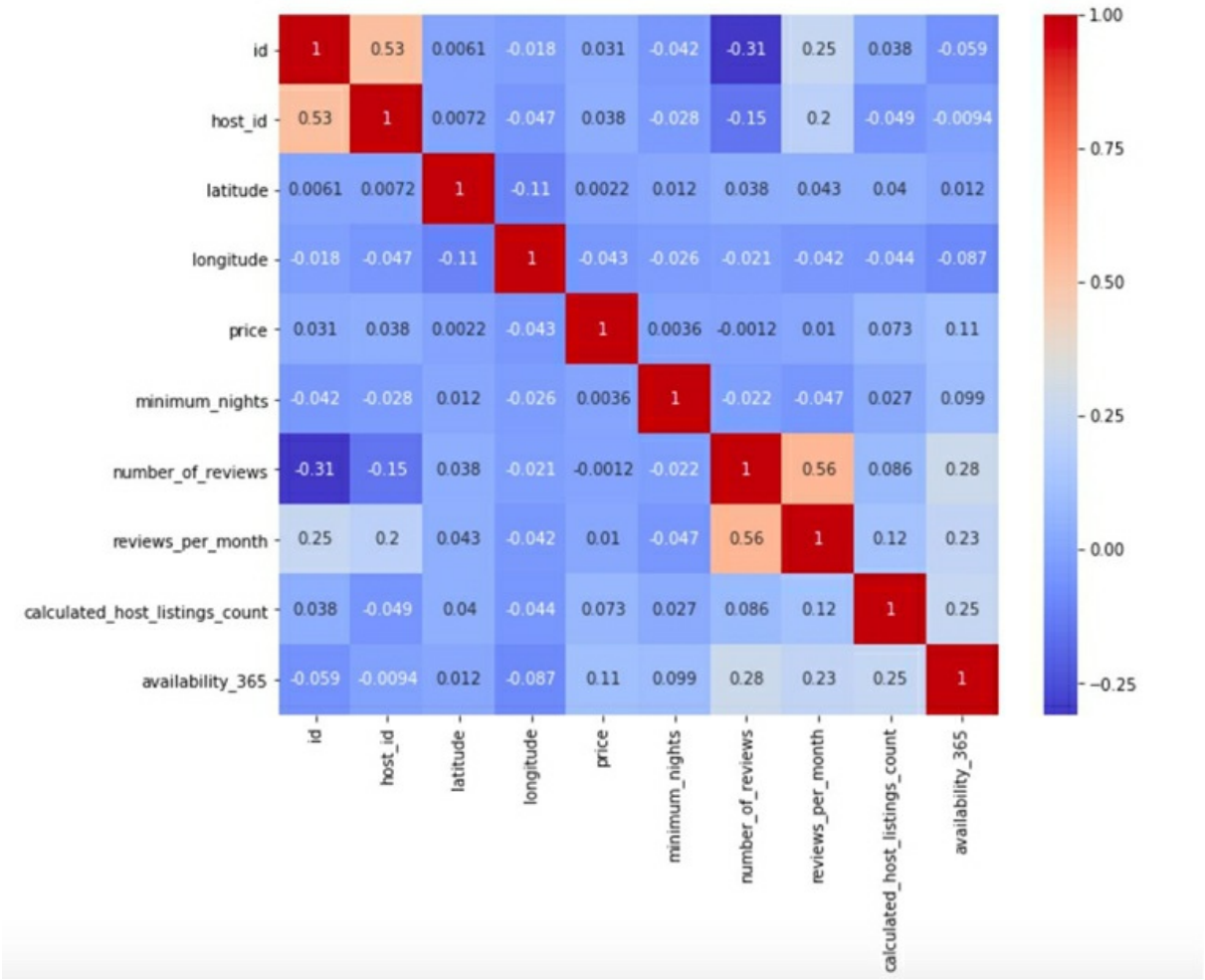
Heatmaps are also useful for inspecting and understanding relationships between variables. The variables are structured as columns and rows on a matrix, with individual values represented as colors on a heat map.

We can build a heatmap in Python using the corr (correlation) function from Pandas and visualize the results using a Seaborn heatmap.

```
df_corr = df.corr()

sns.heatmap(df_corr,annot=True,cmap='coolwarm')
```

```
10 df_corr = df.corr()
11 sns.heatmap(df_corr,annot=True,cmap='coolwarm')
12
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1f25bc18>
```



Example of a heatmap with annotated correlation values

Chapter 6. Combining Different Models for Ensemble Learning

When making important decisions, we generally prefer to collate multiple opinions instead of listening to a single voice or the first person to add their opinion. Similarly, it’s essential to consider and trial more than one algorithm to find your data's best prediction. In advanced machine learning, it can be advantageous to combine models using ensemble modeling, which amalgamates outputs to build a unified prediction model. By combining the output of different models (instead of relying on a single estimate), ensemble modeling helps build a consensus on the data's meaning. Aggregated estimates are also generally more accurate than any one technique. It’s vital, though, for the ensemble models to display variation to avoid mishandling the same errors.

In classification, multiple models are consolidated into a single prediction using a voting system based on frequency or numeric averaging in regression problems. Ensemble models can also be divided into sequential or parallel and homogenous or heterogeneous.

Let’s start by looking at sequential and parallel models. In the former's case, the model’s prediction error is reduced by adding weights to classifiers that previously misclassified data. Gradient boosting and AdaBoost (designed for classification problems) are both examples of sequential models. Conversely, parallel ensemble models work concurrently and reduce error by averaging. Random forests are an example of this technique.

Ensemble models can be generated using a single technique with numerous variations, known as a homogeneous ensemble, or through different techniques, known as a heterogeneous ensemble. An example of a homogeneous ensemble model would be multiple decision trees to form a single prediction (i.e., bagging). Meanwhile, an example of a heterogeneous ensemble would be the usage of k-means clustering or a neural network in collaboration with a decision tree algorithm.

Naturally, it’s crucial to select techniques that complement each other. For instance, neural networks require complete data for analysis, whereas decision trees are competent at handling missing values. Together, these two techniques provide added benefit over a homogeneous model. The neural network accurately predicts the majority of instances where a value is provided. The decision tree ensures no “null” results that would otherwise materialize from missing values using a neural network.

While an ensemble model's performance outperforms a single algorithm in most cases, the degree of model complexity and sophistication can pose a potential drawback. An ensemble model triggers the same trade-off in benefits as a single decision tree and a collection of trees. The transparency and ease of interpretation of, say, decision trees are sacrificed for the accuracy of a more complex algorithm such as random forests, bagging, or boosting. The model's performance will win out in most cases, but interpretability is a crucial factor to consider when choosing the right algorithm(s) for your data.

In terms of selecting a suitable ensemble modeling technique, there are four main methods: bagging, boosting, a bucket of models, and stacking.

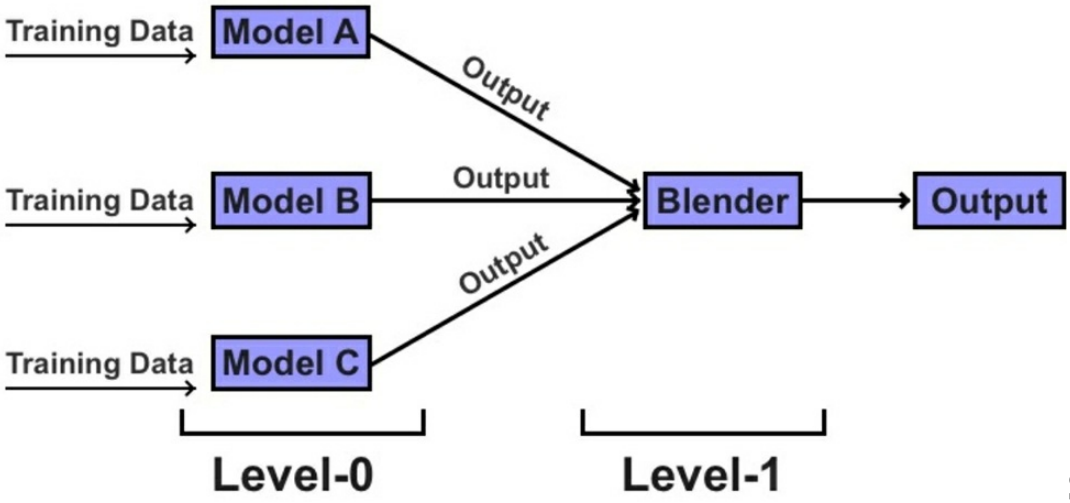
A bucket of models trains multiple different algorithmic models using the same training data as a heterogeneous ensemble technique. It then picks the one that performed most accurately on the test data.

Bagging, as we know, is an example of a parallel model averaging using a homogenous ensemble, which draws upon randomly drawn data and combines predictions to design a unified model.

Boosting is a popular alternative technique that is still a homogenous ensemble but addresses error and data misclassified by the previous iteration to produce a sequential model. Gradient boosting and AdaBoost are both examples of boosting algorithms.

Stacking runs multiple models simultaneously on the data and combines those results to produce a final model. Unlike boosting and bagging, stacking usually combines outputs from different algorithms (heterogenous) rather than altering the same algorithm's hyperparameters (homogenous). Also, rather than assigning equal trust to each model using averaging or voting,

stacking attempts to identify, and emphasize well-performing models. This is achieved by smoothing out the error rate of models at the base level (known as level-0) using a weighting system before pushing those outputs to the level-1 model. They are combined and consolidated into a final prediction.



Stacking algorithm

While this technique is sometimes used in industry, the gains of using a stacking technique are marginal in line with the complexity level, and organizations usually opt for the ease and efficiency of boosting or bagging. However, stacking is a go-to technique for machine learning competitions like the Kaggle Challenges and the Netflix Prize. The Netflix competition, held between 2006 and 2009, offered a prize for a machine learning model that could significantly improve Netflix's content recommender system. From the team BellKor's Pragmatic Chaos, one of the winning techniques adopted linear stacking that blended predictions from hundreds of different models using different algorithms.

Chapter 7. Applying Machine Learning to Sentiments Analysis

Natural Language Processing (NLP) is a widely used field within Artificial Intelligence, which mainly involves the interactions between the human language and the computer. You can find its applications in a large variety of areas such as Sentiment analysis, Spam detecting, POS (Part-Of-Speech) Tagging, Text summarization, Language translation, Chatbots, and so on.

1. How Would you Explain NLP to a Layman? Why Is it Difficult to Implement?

NLP stands for Natural Language Processing, the ability of a computer program to understand the human language. It is an extremely challenging field for obvious reasons. NLP requires a computer to understand what humans speak. But human speech is very often not precise. Humans use slang, pronounce the words differently, and have the context in their sentences, which is very hard for a computer to process correctly.

2. What Is the Use of NLP in Machine Learning?

At present, NLP is based on Deep Learning. Deep Learning algorithms are a subset of Machine Learning, which needs a large amount of data to learn high-level features from data independently. NLP also works on the same approach, uses deep learning techniques to learn human language, and improve upon itself.

3. What Are the Different Steps in Performing Text Classification?

Text classification is an NLP task used to classify text documents into one or more categories. Classifying whether an email is spam or not, analyzing a

person’s sentiments from his post, etc., are text classification problems.

A Text classification pipeline involves the following steps in order:

A. Text cleaning

B. Text annotation to create the features

C. Converting those features into actual predictors

D. Using the predictors to train the model

E. Fine-tune the model to improve its accuracy.

4. What Do you Understand by Keyword Normalization? Why Is it Needed?

Keyword normalization, also known as text normalization, is a crucial step in NLP. It is used to transform the keyword into its canonical form, making it easier to process. It removes stop words such as punctuation marks, words like "a," "an," "the" because these words generally do not carry any weight. After that, it converts the keywords into their standard forms, which improves text matching.

For instance, reducing all words to lower cases, converting all tenses to simple present tense. So, if you have "decoration" in one document and "Decorated" in the other, then both of them would be indexed as "decorate." Now, you can easily apply a text-matching algorithm on these documents, and a query containing the keyword "decorates" would match with both of the documents. Keyword normalization is an excellent means of reducing dimensionality.

5. Tell me about Part-Of-Speech (POS) Tagging.

Part-of-speech tagging is a process of marking the words in a given text as a part of speech, such as nouns, prepositions, adjectives, verbs, etc. It is an extremely challenging task because of its complexity and because the same word could represent a different part of speech in different sentences.

There are generally two techniques used to develop POS tagging algorithms.

The first technique is stochastic, which assumes that each word is known and can have a finite set of tags that are learned during training. The second technique is rule-based tagging, which uses contextual information to tag each word.

6. Have you Heard of the Dependency Parsing Algorithm?

Dependency Parsing algorithm is a grammar-based text parsing technique used to detect noun phrases, noun phrases, subjects, and objects in the text. "Dependency" implies the relations between the words in a sentence. There are various methods to parse a sentence and analyze its grammatical structure. Some of the standard methods include Shift-Reduce and Maximum Spanning Tree.

7. Explain the Vector Space Model and its Use.

Vector Space Model is an algebraic model used to represent an object as a vector of identifiers. Each object (such as a text document) is written as a vector of terms (words) present in it with their weights.

For instance, you have a document "d" with the text "This is an amazing journal for the interview preparation."

The corresponding vector for this document is:

There exist many ways to calculate these weights. They can be as simple as just the frequency (count) of the words in a document. Similarly, any query is also written in the same fashion. The vector operations are used to compare the query with the documents to find the most relevant documents that satisfy the query.

Vector Space Model is used extensively in the fields of Information Retrieval and Indexing. It provides a structure to the unstructured datasets, thereby making it easier to interpret and analyze them.

8. What Do you Mean by Term Frequency and Inverse Document Frequency?

Term Frequency (tf) is the number of times a term occurs in a document divided by the total number of terms in that document.

Inverse Document Frequency (idf) is a measure of how relevant is the term across all the documents. Mathematically, it is logarithmic (total number of documents divided by the number of documents containing the term).

9. Explain Cosine Similarity in a Simple Way.

Cosine similarity captures the similarity between two vectors. As explained in the vector space model, each document and the query is written as a vector of terms.

The cosine is calculated for the query vector with each document, which is the average cosine between two vectors. The resulting cosine value represents the similarity of the document with the given query. If the cosine value is 0, then there is no similarity at all, and if it is 1, then the document is the same as the query.

10. Explain the N-Gram Method.

Simply put, an N-gram is a contiguous sequence of n items in the given text. N-gram method is a probabilistic model used to predict an item in a sequence based on the previous n-1 items. You can choose the items to be either the words, phrases, etc. If n is 1, then it is called 1-gram; for n = 2, it is 2-gram or bigram.

N-grams can be used for approximate matching. Since they convert the sequence of items into a set of n-grams, you can compare one sequence with another by measuring the percentage of common n-grams in both of them.

11. How Many 3-Grams Can Be Generated from this Sentence "I Love New York Style Pizza"?

Breaking the given sentence into 3-grams, you get:

A. I love New

B. love New York

C. New York style

D. York-style pizza

We will use the CountVectorizer package to demonstrate how to use N-Gram with Scikit-Learn.

CountVectorizer converts a collection of text documents to a matrix of token counts.

In our case, there is only one document.

```
from sklearn.feature_extraction.text import CountVectorizer

# N-gram_range specifies the lower and upper boundary on the range of N-gram tokens

# to be extracted. For our example, the range is from 3 to 3.

# We have to specify the token pattern because, by default, CountVectorizer treats single character words as stop words.

vectorizer = CountVectorizer(ngram_range=(3, 3),
                             token pattern = r"(?u)\b\w+\b",
                             lowercase=False)

# Now, let's fit the model with our input text

vectorizer.fit(["I love New York style pizza"])

# This will populate vectorizer's vocabulary_ dictionary with the tokens.

# Let's see the results of this vocabulary

print(vectorizer.vocabulary_.keys())
```

12. Have you Heard of the Bag-Of-Words Model?

The Bag-of-Words model is a widespread technique used in Information Retrieval and Natural Language Processing. It is also known as the Vector Space Model, described in detail in question 6 above. It uses the frequency of occurrence of the words in a document as the feature value.

One of the limitations of this method is that it does not take into account the order of the words in a document, due to which you cannot infer the context of the words. For instance, if you take these two sentences, "Apple has become a trillion-dollar company" and "You should eat an apple every day," the Bag-of-Words model won’t be able to differentiate between Apple as a company and Apple as a fruit. To address this limitation, you can use the N-gram model, which stores the words' spatial information. Bag-of-Words is a particular case of the N-gram method with n=1.

Chapter 8. Conditional or Decisional Statements

These will be an essential part of the code that we work with because they will ensure that your system can respond to the input that the user provides to you. It is hard to predict how a user is going to work with the system. However, you can set up some of the conditions you would like to look at and work from there to develop the way your program will work.

As we can imagine here, it is pretty much impossible for a programmer to create something and guess ahead of time what answers or input the user will provide to the program. And the programmer can't be there watching each use work with the program either, which means that they need to work with the conditional statements. When these are set up correctly, it will ensure that the program will run properly and respond to any information that the user is providing to you.

There are many different types of programs that will respond well to the conditional statements we will discuss in this guide. These are pretty simple to work with, and we will take a look at some of the examples of how you can code with these conditional statements.

We will look at the three main types of conditional statements: the if statement, the if-else statement, and the if statement. Let's take a look at how each of these statements works and use these conditional statements.

The If Statement

As we mentioned, there are three types of conditional statements that we can take a look at. The first one that we need to explore a bit is the if statement. Out of the three that we will spend some time on, the if statement is the most basic. These are not going to be used as much as other options because they often leave a bit desired. However, they are a good springboard for learning what these conditional statements are about and working with them.

With the if statement, the program is set up only to proceed if the user provides us with an input that works with the conditions we set ahead of time. If the input that we get from the user doesn't match our conditions, then the program will just stop, and nothing is going to happen.

As we can see already, there will be some issues with this because we don't want the program to stop with the answer. It should still provide us with some of the basis that we need.

There are a few things that will show up with this code. If you have a user go to the program and state that their age is under 18 years, the program will display the listed message. The user can read this message and end the program right there.

But, things can go wrong if the user puts in that their age is above 18 years. This is true for the user, but because it doesn't meet the conditions you coded in. Thus, the program will see it as false. Like the code is written right now, nothing will happen because it isn't set up to handle this. The user will just see a blank screen any time they enter an over 18 years of age.

The If-Else Statement

Now that we have had some time to look at the simple if statement, it is time for us to move on to the if-else statement. The if the statement is an excellent way to get a bit of practice in coding, there will not be all that many times when we are programming and need to work with this kind of statement.

When your user works with the program, you want to make sure that something shows up on the screen no matter what input they use.

If you use the if statement, like in the example above, and the user puts in an answer (above 18 years), the screen will come back blank using the code that we had from before. This is not something that we want to see, so we need to move on to the if-else statement to see what we can do regardless of what information the user puts into the program.

The if-else statement will provide us with output and ensure that we provide these outputs to the user, regardless of the age or other information we provide to the program. With the example above, if the user comes in and

says that they are 40 years old, then the code will still respond to it.

There are a few options that you can use with this one, but with the idea of the voting option that we talked about with the if statement.

With this option, you add in the else statement, which will cover every age that doesn't fall under 18. This way, if the user does list that as their age, something will still appear on the screen for them. This can provide you with more freedom when working on your code, and you can even add in a few more layers to this. If you want to divide it so that you get four or five age groups and each one gets a different response, you simply need to add more if statements to make that happen. The else statement is at the end to catch everything else.

For example, you can take the code above and ask the user what their favorite color is. You could then have if statements to cover some of the primary colors, such as red, blue, green, yellow, orange, purple, and black.

If the user puts in one of those colors, then the corresponding statement will show up on the screen. The else statement will be added to help catch any other colors that the person may try to use, such as pink or white.

The Elif Statements

The third type of conditional statement that we can work within this process is known as the elif statement. These are going to help us add another level to what we did with the other step. However, they are still going to make sure that the codes we write are as easy as possible.

We can create as many of these elif statements as possible in the code, as long as we add in the else statement at the end. The else statement ensures that we can handle any of the other answers that the user puts in, even those we may not have thought of ahead of time.

When working with the elif statement, it will be similar to giving the user a menu to pick from. You can choose how many of these elif statements you would like to have present in the menu, similar to what is found in many games. Then, the user can pick and choose which one they would like to work with. You can then have a particular action happen, or a confident

statement shows up on the program to meet your needs.

Another thing to notice with the elif statement is that you can add many different options as your code needs. It is possible to make a small menu that just has two or three items in it, or it is possible to expand this out to as many of these as you need to make the code work properly.

The fewer options you work within this one, the easier your code writing will be, so keep that in mind when determining how many options are needed.

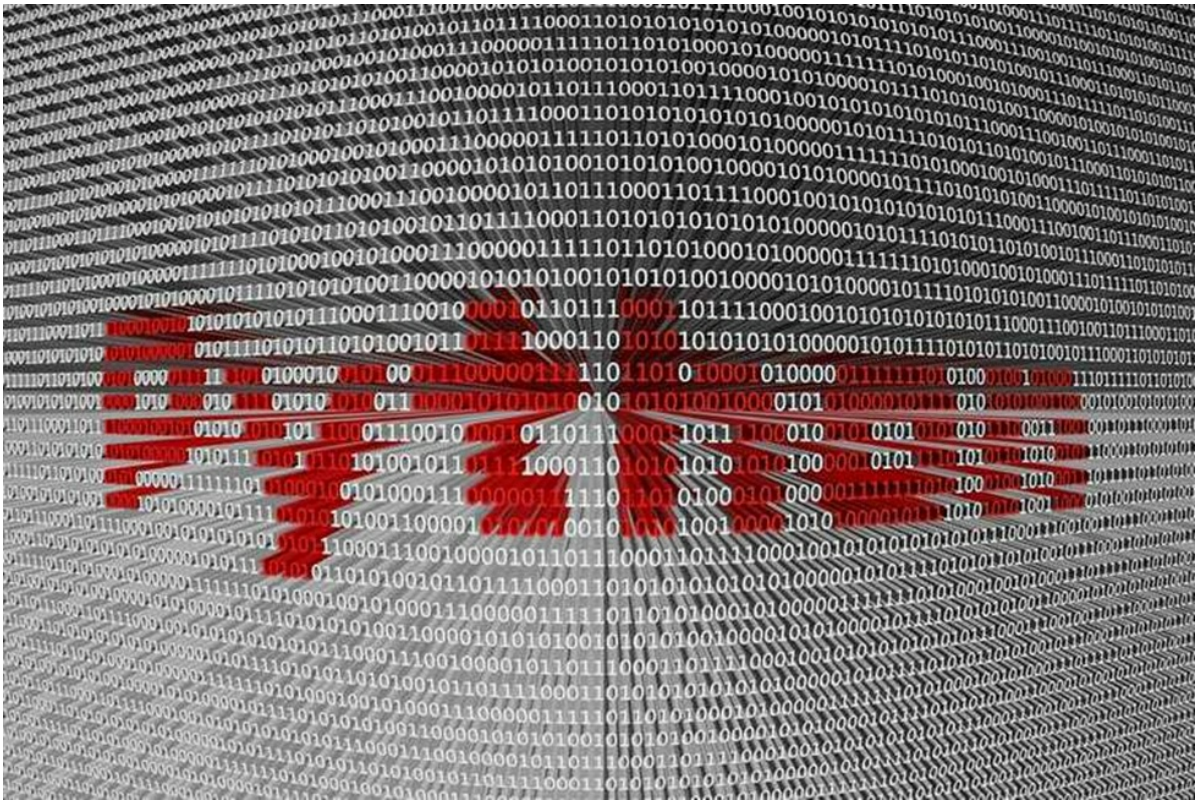
Now that we know a little bit about elif statements and how they work, let's dive in and take a look at a good example of one that you can write out. Open up your compiler and type in the following code:

```
Print("Let's enjoy a Pizza! Ok, let's go inside Pizza hut!")

print("Waiter, Please select Pizza of your choice from the menu")

pizzachoice = int(input("Please enter your choice of Pizza:"))

if pizzachoice == 1:

print('I want to enjoy a pizza Napoletana')

elif pizzachoice == 2:

print('I want to enjoy a pizza rustica')

elif pizzachoice == 3:

print('I want to enjoy a pizza capricciosa')

else:

print("Sorry, I do not want any of the listed pizza's, please bring a Coca Cola for me.")
```

With this option, the user can choose the type of pizza they want to enjoy, but you can use the same syntax for anything you need in your code. If the user enters the number 2 in the code, they will get a pizza rustica. If they don't like any of the options, they tell the program that they want to have something to drink, in this case, a Coca Cola.

Control Flow

The control flow in a program highlights the order of steps of the program execution. In a Python program, control flow is carried out by function calls, conditional statements, and loops. Here, we will deal with if statements, while loops, and for-loops.



Chapter 9. Functions

When you are working with a language like Python, there will be times when you will need to work with something that is known as a function.

These functions will be blocks of reusable code that you will use to get your specific tasks done.

But when you define one of these functions in Python, we need to have a good idea of the two main types of functions that can be used and how they work.

The two types of functions that are available here are known as built-in and user-defined.

The built-in functions are the ones that will come automatically with some of the packages and libraries that are available in Python.

Still, we will spend our time working with the user-defined functions because these are the ones that the developer will create and use for special codes they write.

In Python, though, one thing to remember no matter what kind of function you are working with is that all of them will be treated like objects.

Built-in Functions				
abs()	divmod()	input()	open()	staticmethod()
all()	enumerate()	int()	ord()	str()
any()	eval()	isinstance()	pow()	sum()
basestring()	execfile()	issubclass()	print()	super()
bin()	file()	iter()	property()	tuple()
bool()	filter()	len()	range()	type()
bytearray()	float()	list()	raw_input()	unichr()
callable()	format()	locals()	reduce()	unicode()
chr()	frozenset()	long()	reload()	vars()
classmethod()	getattr()	map()	repr()	xrange()
cmp()	globals()	max()	reversed()	zip()
compile()	hasattr()	memoryview()	round()	__import__()
complex()	hash()	min()	set()	
delattr()	help()	next()	setattr()	
dict()	hex()	object()	slice()	
dir()	id()	oct()	sorted()	

This is good news because it can make it a lot easier to work with these functions than what we may see with some other coding languages.

The user-defined functions will be essential and can expand out some of the work we are doing. But we also need to look at some of the work that we can do with our built-in functions. The list above includes many of the ones that are found inside of the Python language. Take some time to study them and see what they can do to help us get things done.

Why Are User-Defined Functions so Important?

- To keep it simple, a developer will have the option of either writing out some of their functions, known as a user-defined function or going through and borrowing a function from another library, which may not be directly associated with Python. These functions are sometimes going to provide us with a few advantages depending on how and when we would like to use them in the code. When working on these user-defined functions and to gain a better understanding of how they work, some things to remember will be the functions that will be made with reusable code blocks. It is necessary to write them out once, and then you can use them as many times as you need in the code. You can even take that user-defined function and use it in some of your other applications as well.
- These functions can also be handy. You can use them to help with

anything you want, from writing out specific business logic to working on standard utilities. You can also modify them based on your requirements to make the program work properly.

- The code is often going to be friendly for developers, easy to maintain, and well-organized all at once. This means that you can support the approach for modular design.

You can write out these functions independently, and your project's tasks can be distributed for rapid application development if needed. A user-defined function that is thoughtfully and well-defined can help ease the process for the development of an application. Now that we know a little bit more about the basics of a user-defined function, it is time to look at some of the different arguments that can come with these functions before moving on to some of the codes you can use a function.

Options for Function Arguments

Any time you are ready to work with these kinds of functions in your code, you will find that they can work with four types of arguments. These arguments and the meanings behind them are something that will be pre-defined, and the developer is not always going to be able to change them up. Instead, the developer will have the option to use them but follow the rules there. You do get the option to add a bit to the rules to make the functions work the way you want. As we said before, there are four argument types you can work with, and these include:

- Default arguments: In Python, we will find a bit different way to represent the default values and the syntax for your functions' arguments. These default values will be the part that indicates that the function's argument is going to take that value if you don’t have a value for the argument that can pass through the call of the function. The best way to figure out where the default value is will be to look for the equal sign.

- Required argument: The following type of argument is going to be the required arguments. Some kinds of arguments will be mandatory for the function that you are working on. These values need to go

through and be passed in the right order and number when the function is called out, or the code won't be able to run the right way.

- Keyword arguments: These are going to be the argument that will be able to help with the function call inside of Python. These keywords will be the ones that we mention through the function call and some of the values that will go all through this one. These keywords will be mapped with the function argument to identify all of the values, even if you don't keep the same order when the code is called.
- Variable arguments: The last argument that we will take a look at here is the variable number of arguments. This is good for working when you are not sure how many arguments will be necessary for the code you are writing to pass the function. Or you can use this to design your code where any number of arguments can be passed, as long as they have been able to pass any of the requirements in the code that you set.

Writing a Function

Now that we have a little better idea of what these functions are like and some of the argument types available in Python, it is time for us to learn the steps you need to accomplish all of this.

There are going to be four basic steps that we can use to make all of this happen, and it is really up to the programmer how difficult or simple you would like this to be. We will start with some of the basics, and then you can go through and make some adjustments as needed. Some of the steps that we need to take to write out our user-defined functions include:

- Declare your function. You will need to use the "def" keyword and then have the function's name come right after it.
- Write out the arguments. These need to be inside the two parentheses of the function. End this declaration with a colon to keep up with the proper writing protocol in this language.
- Add in the statements that the program is supposed to execute at this time.

- End the function. You can choose whether you would like to do it with a return statement or not.

An example of the syntax that you would use when you want to make one of your user-defined functions includes:

def userDefFunction (arg1, arg2, arg3, …):

program statement1

program statement2

program statement3

Return;

Working with functions can be a great way to ensure that your code will behave the way you would like. Making sure that you get it set up correctly and working through these functions, getting them set up in the manner you would like, can be important. There are many times when the functions will come out and serve some purpose, so taking the time to learn how to use them can be very important to your code's success.

Python Modules

Modules consist of definitions as well as program statements. An illustration is a file name config.py that is considered as a module. The module name would be config. Modules are sued to help break large programs into smaller manageable, organized files, and promote code reusability.

Example

Creating the First module

Def add(x, y):

“This is a program to add two numbers and return the outcome."

outcome=x+y

return outcome

Module Import

Keyword import is used to import.

Example

Import first

The dot operator can help us access a function as long as we know the module's name.

Example

Start IDLE.

Navigate to the File menu and click New Window.

Type the following:

import mine

import mine

import mine

mine.reload(mine)

Dir() built-in Python function

For discovering names contained in a module, we use the dir() inbuilt function.

Syntax

dir(module_name)

Python Package

Files in python hold modules, and directories are stored in packages. A single package in Python holds similar modules. Therefore, different modules should be placed in different Python packages.

Chapter 10. Actual Machine Learning Algorithms

Decision trees are built similarly to support vector machines, meaning they are a category of supervised machine learning algorithms capable of solving both regression and classification problems. They are powerful and used when working with a great deal of data.

You need to learn beyond the barebones basics so that you can process large and complex datasets. Furthermore, decision trees are used in creating random forests, which is arguably the most powerful learning algorithm.

An Overview on Decision Trees

Decision trees are essentially a tool that supports a decision that will influence all the other decisions that will be made. This means that everything from the predicted outcomes to consequences and resource usage will be influenced somehow. Take note that decision trees are usually represented in a graph, which can be described as some kind of chart where the training tests appear as a node. For instance, the node can be the toss of a coin, which can have two different results. Furthermore, branches sprout to represent the results individually, and they also have leaves, which are the class labels. Now you see why this algorithm is called a decision tree. The structure resembles an actual tree. As you probably guessed, random forests are exactly what they sound like. They are collections of decision trees, but enough about them.

Decision trees are one of the most powerful supervised learning methods you can use, especially as a beginner. Unlike other more complex algorithms, they are fairly easy to implement, and they have a lot to offer. A decision tree can perform any common data science task, and the results you obtain at the end of the training process are highly accurate. With that in mind, let’s analyze a few other advantages, as well as disadvantages, to gain a better understanding of their use and implementation.

Let’s begin with the positives:

1. Decision trees are simple in design and, therefore, easy to implement even if you are a beginner without a formal education in data science or machine learning. The concept behind this algorithm can be summarized with a sort of a formula that follows a common type of programming statement: If this, then that, else that. Furthermore, the results you will obtain are very easy to interpret, especially due to the graphic representation.

2. The second advantage is that a decision tree is one of the most efficient methods in exploring, determining the most important variables, and discovering the connection between them. Also, you can build new features easily to gain better measurements and predictions. Don’t forget that data exploration is one of the most important stages in working with data, especially when there are many variables involved. You need to detect the most valuable ones to avoid a time-consuming process, and decision trees excel at this.

3. Another benefit of implementing decision trees is that they are excellent at clearing up some of your data's outliers. Don’t forget that outliers are noise that reduces the accuracy of your predictions. Besides, decision trees aren’t that strongly affected by noise. In many cases, outliers have such a small impact on this algorithm that you can even choose to ignore them if you don’t need to maximize the accuracy scores.

Finally, there’s the fact that decision trees can work with both numerical as well as categorical variables. Remember that some of the algorithms we already discussed can only be used with one data type or the other. On the other hand, decision trees are proven to be versatile and handle a much more varied set of tasks.

As you can see, decision trees are powerful, versatile, and easy to implement, so why should we ever bother using anything else? As usual, nothing is perfect, so let’s discuss the negative side of working with this type of algorithm:

1. One of the biggest issues encountered during a decision tree implementation is overfitting. Please note that this algorithm sometimes creates very complicated decision trees with issues generalizing data due to their complexity. This is known as overfitting, and it is encountered when implementing other learning algorithms as well, however, not to the same degree. Fortunately, this doesn’t mean you should stay away from using decision trees. All you need to do is invest some time to implement certain parameter limitations to reduce overfitting.

2. Decision trees can have issues with continuous variables. When continuous numerical variables are involved, the decision trees lose a certain amount of information. This problem occurs when the variables are categorized. If you aren’t familiar with these variables, a continuous variable can be a value set within a range of numbers. For example, suppose people between ages 18 and 26 are considered of student age. In that case, this numerical range becomes a continuous variable because it can hold any value between the declared minimum and maximum.

3. While some disadvantages can add to additional work in decision trees, the advantages still outweigh them by far.

Classification and Regression Trees

We discussed earlier that decision trees are used for both regression tasks as well as classification tasks. However, this doesn’t mean you implement the same decision trees in both cases. Decision trees need to be divided into classification and regression trees. They handle different problems; however, they are similar since they are both decision trees.

Take note that classification decision trees are implemented when there’s a categorical dependent variable. On the other side, a regression tree is only implemented in a continuous dependent variable. Furthermore, in the case of a classification tree, the training data result is the mode of the total relevant observations. This means that any observations that we cannot define will be predicted based on this value, representing the observation we identify most frequently.

Regression trees, on the other hand, work slightly differently. The value that results from the training stage is not the mode value but the total observations' mean. This way, the unidentified observations are declared with the mean value, which results from the known observations.

Both types of decision trees undergo a binary split, however, going from the top to bottom. This means that the observations in one area will spawn two branches divided inside the predictor space. This is also known as a greedy approach because the learning algorithm seeks the most relevant variable in the split while ignoring the future splits that could lead to an even more powerful and accurate decision tree.

As you can see, there are some differences as well as similarities between the two. However, what you should note from all of this is that the splitting affects the accuracy scores of the decision tree implementation. Decision tree nodes are divided into subnodes, no matter the type of tree. This tree split is performed to lead to a more uniform set of nodes.

Now that you understand the fundamentals behind decision trees, let’s dig a bit deeper into overfitting.

The Overfitting Problem

You learned earlier that overfitting is one of the main problems when working with decision trees, and sometimes it can have a severe impact on the results. Decision trees can lead to a 100% accuracy score for the training set if we do not impose any limits. However, the major downside here is that overfitting creeps when the algorithm seeks to eliminate the training errors, increasing the testing errors. This imbalance, despite the score, leads to terrible prediction accuracy in the result. Why does this happen? In this case, the decision trees grow many branches, and that’s the cause of overfitting. To solve this use, you need to impose limitations on how much the decision tree can develop and how many branches it can spawn. Furthermore, you can also prune the tree to keep it under control, much like how you would do with a real tree to make sure it produces plenty of fruit.

To limit the decision tree's size, you need to determine new parameters during the tree's definition. Let’s analyze these parameters:

1. min_samples_split: The first thing you can do is change this parameter to specify how many observations a node will require to perform the splitting. You can declare anything with a range of one sample to maximum samples. Just keep in mind that to limit the training model from determining the connections that are very common to a particular decision tree, you need to increase the value. In other words, you can limit the decision tree with higher values.

2. min_samples_leaf: This is the parameter you need to tweak to determine how many observations are required by a node, or in other words, a leaf. The overfitting control mechanism works the same way as for the sample split parameter.

3. max_features: Adjust this parameter to control the features that are selected randomly. These features are the ones that are used to perform the best split. To determine the most efficient value, you should calculate the square root of the total features. Just keep in mind that the higher value tends to lead to the overfitting problem we are trying to fix in this case. Therefore, you should experiment with the value you set. Furthermore, not all cases are the same. Sometimes a higher value will work without resulting in overfitting.

4. max_depth: Finally, we have the depth parameter, which consists of the decision tree's depth value. To limit the overfitting problem, however, we are only interested in the maximum depth value. Take note that a high value translates to many splits, therefore a high amount of information. By tweaking this value, you will control how the training model learns the sample's connections.

Modifying these parameters is only one aspect of gaining control of our decision trees to reduce overfitting, boost performance, and accuracy. The

following step after applying these limits is to prune the trees.

Chapter 11. Applications of the Machine Learning Technology

Virtual Personal Assistants

The most popular examples of virtual personal assistance are Siri and Alexa. These systems are capable of providing relevant information using simple voice commands. Machine learning is at the heart of these devices and systems. They collect and define the information generated with every user interaction and use it as training data to learn user preferences and provide an enhanced experience.

Predictions While Driving

Most of the vehicles today utilize GPS navigation services, which collects information such as our current location and driving speed on a centralized server that can generate a map of the current traffic. This helps in managing traffic and reducing congestion. With machine learning, the system can estimate the regions where and the time of the day when traffic jams occur frequently. Machine learning algorithms allow ride-sharing services such as Lyft and Uber to minimize detours on their routes and provide users an upfront estimate of how much the ride will cost.

Video Surveillance

Machines have taken over the monotonous job of monitoring multiple video cameras to ensure the security of premises. Machines can track unusual behavior like standing motionless for an extended period, sleeping on benches, and stumbling. It can then send an alert to the security personnel, who can decide to act on the tip and avoid mishaps. With every iteration of reporting, the surveillance services are improved as the machine learning algorithms learn and improve upon themselves.

Social Media

Social media platforms such as “Facebook,” “Twitter” and “Instagram” are using machine learning algorithms to train the system from user activity and behavior to be able to provide an engaging and enhanced user experience. Some of the examples of the functionalities that are being driven by machine learning algorithms are the “People you may know” feature on Facebook (that collects and learns from user activities such as the profiles they visit often, they’re own profile and their friends to suggest other Facebook users that they can become friends with) and “Similar Pins” feature on Pinterest (that is driven by computer vision Technology working in tandem with machine learning to identify objects in the images of user’s saved “pins” and recommend similar “pins” accordingly).

Email Spam and Malware Filtering

All email clients such as Gmail, Yahoo Mail, and Hotmail use machine learning algorithms to ascertain that the spam filter functionality is continuously updated and cannot be penetrated by spammers and malware. Some of the spam filtering techniques powered by machine learning are Multi-Layered Perceptron and C 4.5 decision tree induction.

Online Customer Service

Nowadays, most e-commerce sites allow users to chat with a customer service representative, usually supported by a Chatbot instead of a live executive. These bots use machine learning technology to understand user inquiries and extract information from the website to resolve customer issues. With every interaction, Chatbots become smarter and more humanlike.

Refinement of Search Engine Results

Search engines such as “Google,” “Yahoo,” and “Bing” use machine learning algorithms to provide improved search results pertinent with the user-provided keywords. For every search result, the algorithm observes and learns from user activity such as opening suggested links, the order in which the opened link was displayed, and time spent on the opened link. This helps

the search engine understand which search results are more optimal and any further modifications to improve the search results.

Product Recommendations

The product recommendation feature has now become the heart and soul of the online shopping experience—machine learning algorithms, combined with artificial intelligence, fuel the product recommendation functionality. The system observes and learns from consumer activity and behavior such as past purchases, wish lists, recently viewed items, and liked or added to cart items.

Online Fraud Detection

Financial institutions rely heavily on machine learning algorithms and artificial intelligence to secure cyberspace by tracking potentially fraudulent monetary transactions online. For example, PayPal is using Machine learning algorithms to prevent money laundering through its platform. They are using artificial intelligence tools in combination with Machine learning algorithms to analyze millions of transactions and discriminate between legitimate and illegitimate transactions between the buyer and the seller. With every transaction, the system learns which transactions are legitimate and which transactions could be potentially fraudulent.

Predictive Analytics

As per SAS, the prescient examination is the "utilization of information, accurate calculations, and AI methods to extricate the probability of future results dependent on verifiable information. The objective is to go past, realizing what has ended up giving the best appraisal of what will occur after on." Today, organizations are burrowing through their past with an eye on the future. This is where human-made consciousness for promoting becomes an integral factor, using proactive examination innovation. The prescient examination's accomplishment is straightforwardly relative to the nature of large information gathered by the organization.

Here is a portion of the broadly utilized prescient examination applications for showcasing:

Prescient Analysis for Customer Behavior

For the modern goliaths like "Amazon," "Apple," and "Netflix," examining client exercises and conduct is essential to their everyday activities. More modest organizations are progressively accepting their function to actualize prescient investigation in their plan of action. The advancement of an altered set-up of prescient models for an organization isn't just capital-concentrated yet requires general labor and time. Showcasing organizations like "AgilOne" offer generally straightforward prescient model sorts with wide materialness across modern areas. They have distinguished three fundamental sorts of prescient models to dissect client conduct, which are:

"Inclination models" – These models are utilized to produce "valid or exact" expectations for client conduct. Probably the most well-known penchant models include: "prescient lifetime esteem," "inclination to purchase," "affinity to turn," "affinity to change over," "probability of commitment," and "inclination to withdraw."

"Bunch models" – These models are utilized to separate and gather clients dependent on shared characteristics, such as sex, age, buy history, and socioeconomics. The absolute most basic group models incorporate "item-based or class base bunching," "conduct customs grouping," and "brand based bunching."

"Communitarian separating" – These models are utilized to create items, administrations, and proposals just as to suggested notices dependent on earlier client exercises and practices. Probably the most widely recognized community sifting models incorporate "upsell," "strategically pitch," and "after sell" proposals.

Organizations' main apparatus to execute prescient examination on client conduct is "relapse investigation," which permits the organization to build up relationships between's offer of a specific item and the particular ascribes showed by the buying client. This is accomplished by utilizing "relapse coefficients," which are numeric qualities portraying how much the client's

conduct is influenced by various factors and building up a "probability score" for the item's future offer.

Capability and Prioritization of Leads

There are three introductory classes utilized in business-to-business or B2B prescient examination promoting to qualify and organize planned clients or "leads."

These classifications are:

- "Predictive scoring" is utilized to organize forthcoming clients based on their probability to make a real buy
- "Identification models" are utilized to distinguish and get new imminent clients dependent on properties imparted to the organization's current clients.
- "Automated division" is utilized to isolate and characterize planned clients dependent on shared characteristics to be focused on the same customized advertising techniques and missions.

The prescient examination innovation needs a huge volume of deal information that fills in as a structure square and preparing material to increment the prescient models' exactness and proficiency. Little physical organizations can't bear to grow their figuring assets; consequently, they can't proficiently gather client conduct information from their in-store deals. This converts into a serious edge for the bigger organizations with a further developed registering framework, which fuels bigger organizations' pointless development in contrast with independent ventures.

Distinguishing Proof of Current Market Trends

Organizations can utilize "information representation" devices that permit business heads and administrators to assemble experiences on the organization's present status, basically by picturing their current client conduct information on a "report or dashboard." These dashboard reports

will, in general, rouse and create client conduct driven activities. For instance, an organization can distinguish the basic client requests pattern in explicit areas with information representation devices and likewise plan to stock their stock for only stores. Similar data can uncover the best items and administrations for the organization to be dispatched depending on the current market drifts that can trick the client requests. The market pattern bits of knowledge can likewise be applied to expand its effectiveness gracefully chain the executives model.

Client Segmentation and Targeting

One of the least difficult and profoundly successful methods of streamlining an item offered to accomplish a fast turnaround on the organization's quantifiable profit is the capacity to target "right clients" with the suitable item offers at the "perfect time." This additionally turns out to be the most well-known and broadly utilized utilization of prescient investigation in the realm of advertising. As indicated by an exploration study directed by the "Aberdeen Group," organizations utilizing prescient investigation in their showcasing methodologies are multiple times bound to distinguish "high worth clients effectively." This is the place where the nature of the organization's current informational index comes first. The energetically prescribed practice utilizes chronicled buyer conduct information of every current client, investigates it to portion, and targets clients with comparable buying credits with a customized proposal and promoting efforts.

The absolute most basic prescient examination models utilized in this application are "liking investigation," "agitate examination," and "reaction demonstrating." Using these applications, organizations can assemble understanding, for example, "if consolidating advanced and print memberships of their item contributions or list is a smart thought" or "whether their item or administration will be more effective whenever offered as a month to month membership model or one-time buy charge." One of the major deals and showcasing stage organizations is "Salesforce," which offers a cloud-based stage that organizations can utilize to create client profiles due to the information gathered from free sources, including client relationships, the executives (CRM) applications, and other organization applications. By specifically and carefully adding inputted information to this stage,

organizations can consistently follow their client conduct to continuously build up a client social model to take care of the organization's emotional cycle continuously and over the long haul.

Advancement of Marketing Strategies

Another use of prescient investigation and showcasing is giving admittance to an assortment of client-related information, such as information gathered from online media stages and organizations' inward organized information. The client conduct model would then be produced by examining all accessible information and applying "social scoring." All the organizations across various mechanical areas must adjust to changing or developing client conduct through multiplying promoting mediums or channels. For instance, organizations can utilize any of the proactive investigation models portrayed above to foresee if their arranged advertising effort would accomplish the online media stages or their versatile applications.

Organizations can utilize the prescient examination model to comprehend how their clients are interfacing with their items or administrations, in light of their sentiments or feelings shared on the online media stages concerning a specific theme.

Chapter 12. Data Mining and Applications

What’s the point of ads? They’re on our monitors, TV screens, smartphone displays, inside our favorite radio broadcasts, and mailboxes. No matter where we turn, we’ll find ads constantly hawking something we’re not interested in. Those ads represent the traditional shotgun approach where companies simply propel as many as they can in our general direction and hope at least one hits. As we can imagine, this kind of marketing costs a lot. Still, companies don’t know any better and keep pumping money into wacky, bizarre, and embarrassing ads, hoping anything works. Thanks to Machine Learning, we might be nearing a future where computers produce dirt-cheap ads that are scarily tailored to our behavior and applied at the exact moment when they’ll have the strongest effect. We might already be living in one such future.

One thing about consumer behavior is that most purchases are made automatically, but major life events can break these habits and put us on the cusp of trying new things. This means Fig Newtons ads aren’t necessarily aimed at people who’d never try Fig Newtons but at those who like sweets and might try something different because they’re undergoing a major life event, such as divorce, car purchase, or pregnancy. How does the advertising company know which person is which? Enter data mining, harvesting as much data about people to have computers try to predict their behavior, desires, and motivations to target them with just the right kind of ad at just the right moment. Of course, ads would never work for us, but machines can learn to be persuasive.

One thing to note here is that data mining processes are going to be used to help us build up Machine Learning models. These models that rely on Machine Learning can power up applications, including the recommendation programs found on many websites and the technology that can keep search engines running.

How Does Data Mining Work?

So, why is data mining such an important process to focus on? You will see the staggering numbers when it comes to the volume of data that is produced is doubling every two years. Just by looking at unstructured data on its own, but just because we have more of this information doesn't mean that we have more knowledge all of the time. With the help of data mining, you can do some of the following tasks:

- Sift through all of the noise, whether repetitive or chaotic, is found in your data.
- You can better understand what is relevant in all of that information and then make good use of the information to assess what outcomes are the most likely for your needs.
- It can help you accelerate the pace of making decisions informed and driven by data and more likely to help your business thrive and grow.

Now that we have that figured out, it is time to look at how all of this data mining will work. We will not grab the data, these trends will show up with us having to do no more work on them, and this is where we will be able to work with data mining. Data mining is a great way for us to explore and analyze a large amount of information to find all of the insights, trends, and patterns that we can use out of that information.

For example, we can work with data mining to learn more about the opinions and the users' sentiment, help us learn how to properly detect fraud, help out with risk management, filter the spam out of email, and even with marketing. All of these are going to be important to many different kinds of companies, and when you use them properly, you will find that they are going to ensure that you can better serve your customers over time.

There are five basic steps that we will see when it comes to working with data mining. In the first step, the company will spend some time collecting the data they want to use, and then they will make sure that all of this will be loaded up properly to their data warehouse. When this is all done, the company can then store and manage the data. Sometimes, this is done on the

company's in-house servers, and other times it is going to be sent to the cloud.

When we go through with this, the management teams, IT professionals, and even business analysts will gain access to this data. Then they can determine the way that they would like to organize all of this information. We can then work with application software to sort out the data based on the results that the user is going to put in. In the last step, our end-user will present their findings and all of that information in a certain format that makes the most sense, that will be easy for those in charge of making decisions to read through and understand.

While we are on this topic, we need to work on data warehousing and mining software. The different programs that you decide to use with data mining will be responsible for analyzing the patterns and the relationships that we can find in the data. All of this is going to be done based on the requests that the user sends out. A company may use this software to help them create some new classes on that information.

We can go through and illustrate this point a bit more, as well. Imagine that we are a restaurant that would like to work with all of the data mining steps to determine the right times to offer some specials. The restaurant would be able to do this by looking at all of the information they have been able to collect on the specials, and then see how the specials do at different times of the day and on different days of the week. They can then create classes based on when the customers visit and what the customer is most likely to order when they come to the restaurant to eat.

We can take this to the following level as well. In some cases, a data miner will find clusters of information based on a logical relationship, or they may take some time to see if there are sequential patterns and associations that they can draw some conclusions to learn more about their customers in the long run.

Warehousing is going to be another important part that we see in the data mining process. This is a pretty simple process to work with, and it is going to include when a company can centralize their data into one database or one program, rather than spreading out the information in more than one place. With the warehouse for data, an organization can spinoff some of the data

segments for the right users to analyze regularly and for the specific users to gain access to when they need it.

However, there are also times when we will see that the analyst will take the work on a different course during this process. For example, the analyst may choose to start with some of the most useful data, and then they will be able to create their warehouse for the data based on the specifications there. No matter how a business wants to organize their data, they will use it to help support some of the decision processes that the company's management is going to make.

With this in mind, we also need to take some time to explore data mining examples along the way. A good example of this is grocery stores. Many of the supermarkets that we visit regularly give away free loyalty cards to customers. These are beneficial to the customers because it provides them with access to reduced prices and other special deals that non-members at that store will not be able to get.

This is a great way for both parties to win. The customer will enjoy that they can potentially save money so that they will sign up for it. The store will enjoy that they get a chance to learn more about the customers, set prices to bring in more people, and make them the most money possible.

Here, we need to keep in mind that there are a few concerns that data mining can bring up for the customer and a company. Some customers are concerned about this data mining process because they worry about the company not being ethical with their use. It could even be an issue with a legitimate and honest company because the sampling they use could be wrong, and then they will use the wrong kind of information to make their decisions.

Most companies need to take some caution when they decide to work with the data mining process to ensure that they will reach their customers better and do well in their industry through some useful insights and more that they can learn along the way. You need to focus on when learning what patterns and insights are found in all of that data.

All of this data is going to be important when it comes to working in the process of data science. But we have to make sure that we understand how this data is supposed to work and what is found inside of all that data. When

we can learn all of this, we will find that it is easier than we may think to handle the data and work for our needs.

Unbalanced Data Set

Although imbalanced data is a common problem with datasets, there is no universal technique for dealing with this issue. Generally, when classifiers are fed imbalanced data, the classification output will be biased, resulting in always predicting the majority class and incorrectly classifying the minority class. Therefore, we need to detect when the output is biased and deal with this issue, to improve its accuracy. We will over-sample the minority class by employing the Synthetic Minority Over-Sampling Technique (SMOTE) and the stratified K-Fold cross-validation method for dealing with the class imbalance.

Conclusion

Machine learning is an exciting and rapidly evolving field. While mastery of the subject can involve many years of study, it is possible to get started quickly by gaining some basic familiarity with machine learning methods and goals.

Despite the mysterious aura surrounding the field, many of the machine learning methods are relatively simple mathematical tools that have been around for centuries. It is just now that they are being applied to the massive amounts of data, the so-called big data, that is being collected by companies and other large organizations.

Python is an excellent tool to use for learning about–machine learning. Python is a very simple programming language that most people can pick up rather quickly. Libraries have been developed for python that is specifically designed for machine learning. So it is easy for a developer to play around with the tools and solve simple machine learning problems.

The way to go forward is actually to practice and study more. Begin by going through any exercises that you can find that entail covering all of the major algorithms used in machine learning. Using both supervised and unsupervised learning is important, as anyone who wants to understand machine learning needs to become intimately familiar with both. You should also practice by using many of the standard algorithms like linear regression and k-nearest neighbors.

Something that I would suggest is to avoid getting trapped into only using generated test data. To enhance your learning and development, get a hold of real-world data sets that you can run your algorithms on to gain an even greater familiarity with the practice of data science.

Many people who are new to the concept of machine learning ask what specific educational credentials they need to get into the field. While there are some general guidelines, the truth is there are no specific rules. We can begin by saying that in all likelihood, anyone who is involved in a scientific or

technical field of study would be in a position to get involved in machine learning. That certainly applies to electrical or computer engineers.

However, some people who might be better placed to get into machine learning are mathematicians experts in statistics and probability. Some crossover knowledge can be helpful, but in some ways, machine learning is a statistical field when it comes down to a data scientist's day-to-day practice. Certainly, a high level of knowledge of statistics and probability is helpful.

Since it is considered a crossover discipline background in computer science can be helpful. The ideal candidate would have a substantial computer science background that has also demonstrated a high-level education in statistics. The more advanced your education, the more deeply you can go into the field, including doing AI research and designing more advanced systems. If you are playing around with some models, you will not be designing machine learning systems for use in some new robotic systems. That will require advanced education in computer science.

However, there are varying roles and levels of machine learning. Those who study computer systems in business school provided that they have a good understanding of statistics will be well-suited for doing machine learning tasks as a data scientist at many companies. Simply analyzing customer data or internal company data for trends and patterns is not something that requires a deep understanding of artificial intelligence. Your role is to use machine learning tools available to extract the kind of information useful for the enterprise.

So, machine learning and data science are fields that have a wide range of complexity and application. There is virtually some level of expertise suitable for many different levels and types of education, background, and taste. It is a growing field for the future.

We have learned what machine learning is and how it is applied today by businesses to many different tasks. We learned that there is supervised, unsupervised learning and how they are different. We also learned the issues that might crop up with various tradeoffs in machine learning.

We also learned many of the major algorithms used in machine learning, including regression methods, k-nearest neighbor methods, and decision

trees. A large part of building a solid and reliable machine learning system is selecting the most appropriate training data sets and the best algorithm for a given situation. This, in part, will be determined by your experience, and the more experience that you get practicing machine learning, the better you are going to be when it comes to selecting the right algorithms for a given problem.

We also saw how python could be used to implement some of the most common machine learning tasks. We used python for regressing, k-nearest neighbors, and other classification methods. We looked at the TensorFlow library, the Sciikit learn library, and the Numpy library. We also learned about Keras and saw how to build a neural network. The power of Numpy lies behind many of the tools used to build machine learning models with python.

So where to go from here? The first step is to keep learning. You should keep practicing by building more models and using different tools to build your models. However, there is more to machine learning than simply playing around with tools. You should read as many journals as you can and watch videos from reputable sources so that you can learn the theory and fundamentals that lie behind the concept of machine learning.

If you go further than this, it will largely depend on your current situation and your future goals. If you are already a working professional, you might not need to go to school and get a computer science degree. You might be learning the tools for the sake of practical application at your current job. If that is the case, then practice along with self-study is the best path forward for you, although of course, if you are willing and able to return to school to get an in-depth education on the subject, that is always an option.

For those who are just getting exposed to the field and looking to it to pick a career path, getting a college degree in computer science or a related subject is probably the best way forward for you, especially if you hope to attain employment. Data science and machine learning are not likely to be fields where too many people can get employment without some college degree in a related field. If possible, find a school that will let you concentrate on artificial intelligence and machine learning.

I would also advise taking many math classes that are focused on statistics

and probability. Some “business acumen” is often advised, so it can’t hurt to take some management classes as well. This is recommended even though many technical types are not that enthusiastic about business school. You are not going there to become an MBA. Still, you should get some idea of business operations at a large corporation and learn about many business concepts like business intelligence, predictive analytics, and data mining, since these are useful concepts for corporations. They prefer people who understand this to join their team, ready to hit the ground running.

Computer engineering is a related field that can also be pursued, and you can even consider mechanical engineering. That might not come to mind right away, but remember that there is a lot of research in robotics in mechanical engineering. But remember that college is nothing more than an entry ticket. Machine learning is a very practical field, and many of the tools described are going to be used in the real world.

I hope that has stimulated your interest in machine learning and that it will help propel you to continue your education and development in this exciting area.

PART II

Introduction

Python is among the most popular computer language programming tool initially created and designed by Guido Van Rossum in the late 1980s. Since its introduction into the computing world, Python has undergone multiple modifications and improvements, becoming among the leading programming languages used by developers. The tool is dynamically typed, object-oriented, multi-paradigm, and imperative. It is used across different operating systems, including Windows, Linux, Android, macOS, and iOS devices. It is also compatible with both bit 32 and bit 64 gadgets of phones, laptops, and desktops.

Despite comprising of several areas essential for programmers, Python is easy to learn, especially for beginners with minimal computer programming knowledge. Unlike most programming languages, Python accompanies an easy to use syntaxes where first time users can readily practice and become a pro within a few weeks. However, the programming processes may vary depending on the motive of the learner in programming. Despite accompanying multiple vocabularies and sometimes sophisticated tutorials for learning different programming techniques, engaging with Python is worth developing excellent programs.

Features of Python Programming

Simple Language

Most programming languages have complicated and lengthy coding languages, which may become cumbersome to beginners. Long and challenging languages may become hard to learn and remember, therefore hindering amateurs' learning abilities. Python accompanies very simple and fantastic syntax, making beginners read and write programs without complications readily. Compared to Java and C++, Python enables you to work with ease while focusing on the outcomes.

Portability

With its ability to run in any operating system, Python allows for portability where you can readily transfer your codes and the general program from one device to the other without affecting your progress. This programming tool is quite useful for developers who change devices or transfer data from one platform to another. You can, therefore, run your program in the new machine seamlessly with little alliteration. Besides, Python allows the continuation of your application to your primary system and effectively run as intended.

Standard Libraries

Today, all programming languages consists of libraries where you can quickly select a program, make modifications are necessary, and execute your codes. Some of these libraries may have limited coding lines, which will, therefore, require you to write your program. On the other hand, Python comes with an extensive standard library that comprises all your programming needs. For example, it consists of the MySQLdb library, which allows you to connect to the MySQL database without creating a pathway. As such, Python becomes among the leading programming tools to be commercially used when dealing with thousands of data as you can quickly retrieve and run with ease.

Free Open-Sources

Python also offers a fantastic free and open-source where you can use the tool in different areas, such as commercial use. Unlike other programming tools, a developer can choose to make changes to the program or select the desirable dataset to suit the field at hand, mainly in the Python source code. While being used across several areas in the computing community, Python has experienced a constant increase in usage, becoming more simple benefitting beginners.

Downloading and Installing Python

Like most computer software, Python can be downloaded, run, and installed in a system for it to function with ease. However, this tool may become challenging during download or updating, depending on the operating system. Some systems such as macOS and Linux typically accompany a preinstalled Python version, which is mostly outdated. These versions of Python will hence require an update, which usually uses unique techniques. On the contrary, other operating systems such as Windows and Android devices require a user to visit the Python homepage or other relevant websites, download and install the software.

Python Development and Application

Python development is usually undertaken by the Python Enhancement Proposal (PEP), which has led to creating the most advanced and latest version. PEP has enhanced the features, Python documentation, and the creation of bug fixes essential for eliminating problems arising during programming. Besides, it has managed to design modern coding processes and extend standard libraries to suit all developer needs during the creation of programs. In most cases, PEP collects information from developers utilizing Python and develops solutions to major problems raised.

When Python was first released in the 1980s, it accompanied multiple benefits but with numerous faults within the tool. Over the years, the Python Software Foundation has made significant modifications indicating differences between Python 1.0 and Python 3.7 used today. Python has henceforth gained popularity over time and applied in various areas in the computing community. For instance, the programming software has been used to create Web apps in different websites such as Instagram, Mozilla, and Reddit. Other applications include the computation of both scientific and numerical values and the development of software prototypes. Due to it’s

easy to use language, Python is widely used in educating children and beginners interested in learning computer language skills.

Python Variables

Python variables are named sections used to store codes in the system memory used mainly to develop programs. Variables are critical in Python, especially for programmers who create complex programs in need of multiple code values. Unlike other programming software like Java and C++, Python doesn't demand variable declaration as they instantly change after being named. Python variables, therefore, are memory reserves used to store values fed to the program when needed. The data saved usually varies depending on the data type; for instance, they may be stored in Numbers, Lists, Tuple, or as Dictionary.

Lists include ordered and changeable data written in the form of "my computer" with double-quotes. You can access values within the list using index numbers, which are written up to negative integers. Dictionaries entail indexed and changeable variables but remain unordered and written with single curly quotes. Accessing values in dictionaries consist of inputting a keyword in parentheses, which also helps other functions such as looping, making changes, etc. Numbers are of three forms int, float and complex representing different number value stored while tuple consists of data values which are ordered but remain unchangeable.

Naming Variables in Python

The naming of variables, especially in Python, is useful as it makes these storage units easily identified by a different programmer. Naming is smooth, unlike other programming tools, as beginners can assign a name to a given variable. However, assigning titles in any computer language programming process follows specific rules to ensure that the names are practical and easily recognized. Some names may lack a desirable representation of what is included in a given variable, thus causing confusion between programmers. Some of these rules are:

- Titles must be of singles words with no spaces between letters or numerical
- First characters must never be a number
- Never use reserved words as variable names
- The name must consist solely of letters and numbers with underscores acting as spacers
- The name must begin with a lowercase letter

In case you assign names that do not follow these rules, the system will reject the name as it is case sensitive. Customarily, the system may act as a guide when naming your variable as it readily notifies you where the mistake has been made. There are some situations where a developer may choose to assign multiple names to one variable. That is, writing two or more words to describe what is included in the variable. In this case, you are eligible to create your name but following either one of the methods used.

The Pascal case method is one way you can use it as it involves the first and subsequent words to be capitalized for readability enhancement. An example of Pascal's case is Python Programming Language. Another method is the Camel case and is where the second and subsequent words are capitalized. An example is Python Programming Language. Lastly, the Snake case is another method, and this uses underscores as spacers when creating your variable name. For instance, python_Programming_Language. All the three modes of multi-name given to Python variables are correct, and you can choose from them while assigning your variable a title.

Types of Data Variables

Int

This is a number data variable stored in 16-bit values and ranges between -32,768 and 32,767 in Python but depends on other programming tools. Int stores up to 2's complement math, suggesting that it can provide reserves for negative numbers. Therefore, int has a higher probability of providing adequate storage units of quite smaller amounts. When dealing with arithmetic variables, then int plays a significant role in feeding your program

with the intended data.

Char

Char are data variables used to store data codes expressed in literal values and written with single quotes such as ‘A.’ The values are also numerical but with direct visibility to the codes used in a given program. Char, therefore, makes the performance of arithmetic functions quite useful as the data is usually stored in 8-bit, but those with higher memory usage being stored in bytes. Chars are typically the smaller storage units of bytes.

Bytes

Bytes are much higher data storage units essential for storing values with higher memory usage and those which cannot be stored in chars. More so, bytes are also used to store 8-bit unsigned numbers, which are from 0-255. Bytes and chars play a similar role as data type storage reserves of numbers but vary in the size of values stored in each section.

Strings

This is another form of a data variable that creates a series of char data types or data stored in a chain. The syntax used has several declarations before the values are marked for use as strings comprise arrays of chars. Strings are typically showed with double quotes and may store a large number of values within one chain. The chars can also be broken down to form other chains as though they would require many declarations when retrieving the needed data to create a program.

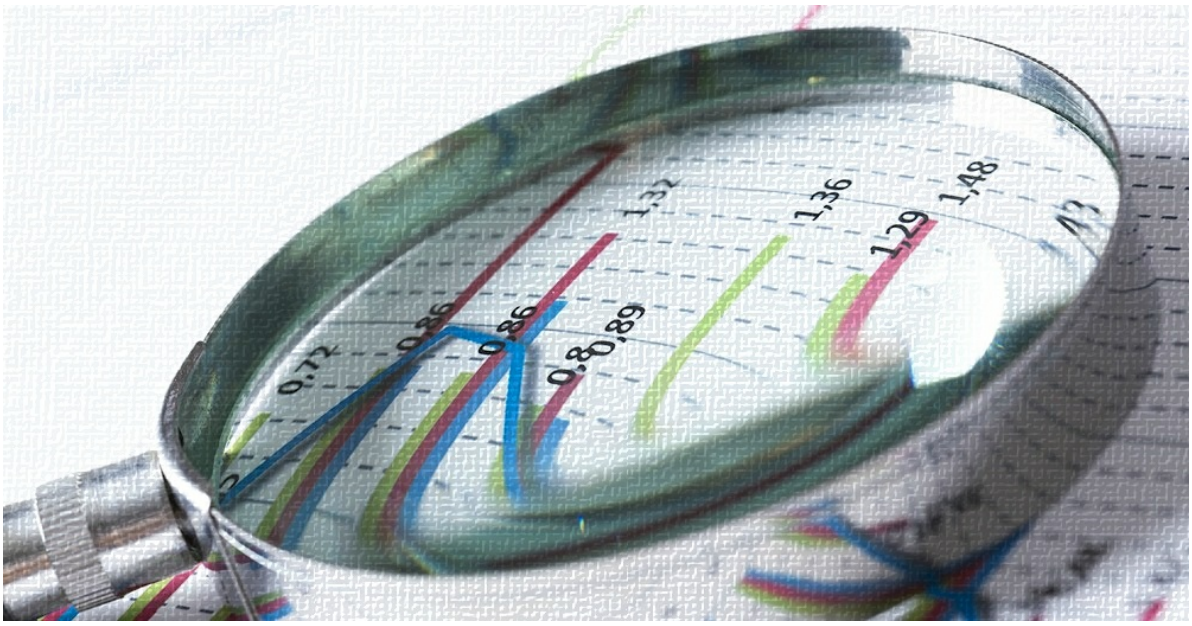
Python Debugging

Debugging is the process or technique used to detect and eliminate problems that arise during a program’s writing and execution. Since its incorporation in the 1940s, computer debugging has become one of the techniques used to prevent errors, bugs, and mistakes arising during programming processes. The direct opposite of the term is anti-debugging, which entails reversing, detecting, and removing such errors with tools such as modified codes, API-

based, and timing and latency.

In Python, the software also includes debugging but primarily depends on a Python interpreter to recognize and eliminate problems. In some cases, Python debugging is quite effective and enables programmers to create programs after every breakpoint. When writing codes, you may continually input your codes without recognizing errors, bugs, or even typos that may affect your outcome. Therefore, debuggers tend to indicate these problems and may either provide solutions instantly or take a breaking point for you to correct it.

Chapter 1. About Data Analysis



Companies have spent a lot of time looking at data analysis and what it has been able to do for them. Data is all around us, and it seems like tons of new information is available for us to work with regularly each day. Whether you are a business trying to learn more about your industry and your customers, or just an individual who has a question about a certain topic, you will be able to find a wealth of information to help you get started.

Many companies have gotten into a habit of gathering up data and learning how to make it work according to their needs. They have found that there are many insights and predictions within data to make sure that it is going to help them out in the future. If the data is used properly, and we can gain a good handle of that data, it can help our business become more successful.

Once you have gathered the data, there is going to be some work to do. Just

because you can gather up all of that data doesn't mean that you will see what patterns are inside. This is where data analysis is going to come into play to help us see some results as well. This process is meant to ensure that we fully understand what is inside of our data and can make it easier to use all of that raw data to make some informed and smart business decisions.

Data analysis, to make this a bit further, will be a practice where we can take some of the raw data that our business has been collecting and then organize and order it to ensure that it can be useful.

We will find that with all of these methods, it is easier for us to work with data analysis because we can make some of the adaptations that are needed to the process to ensure it works for our own needs, no matter what industry we are working in, or what our main question is in the beginning.

The one thing that we need to be careful about when we are working with data analysis is to be careful about how we manipulate the data that we have. It is really easy for us to go through and manipulate the data wrongly during the analysis phase and then push certain conclusions or agendas that are not there. This is why we need to pay some close attention to when the data analysis is presented to us and think critically about the data and the conclusions that we were able to get out of it.

If you are worried about a source that is being done, and if you are not sure that you can complete this kind of analysis without some biases, it is important to find someone else to work on it or choose a different source. There are plenty of data out there, and it can help your business to see some results, but you have to be careful about these biases, or they will lead us to the wrong decisions in the end if we are not careful.

Besides, you will find that during the data analysis, the raw data you will work with can take on various forms. This can include things like observations, survey responses, and measurements, to name a few. The sources that you use for this kind of raw data will vary based on what you are hoping to get out of it, what your main question is all about, and more.

In its raw form, the data that we are gathering will be very useful to work with, but you may find that it is a bit overwhelming to work with. This is a problem that many companies will have when they work with data analysis

and something that you will have to spend some time exploring and learning more about.

Over the time that you spend on data analysis and all of the steps that come with the process, the raw data will be ordered in a manner that makes it as useful to you as possible. For example, we may send out a survey and tally up the results that we get. This will be done because it helps us see at a glance how many people decided to answer the survey at all and how people were willing to respond to some of the specific questions that were on that survey.

In organizing the data, a trend is likely going to emerge, sometimes even more than one. Besides, we will be able to take some time to highlight these trends, usually in the write-up that is being done on the data. This needs to be highlighted because it ensures that the person reading that information is going to take note.

There are plenty of places that we are going to see this. For example, in a casual kind of survey that we may try to do, you may want to figure out the

preferences of what ice cream flavors men and women like the most. In this survey, maybe we find out that women and men will express a fondness for chocolate. Depending on who is using this information and what they are hoping to get out of that information, it could be something that the researcher is going to find very interesting.

Modeling the data found out of the survey or another method of data analysis using mathematics and some of the other tools out there can sometimes exaggerate the points of interest, such as the ice cream preferences. This will make it so much easier for anyone looking over the data, especially the researcher, to see what is going on there.

In addition to looking at all of the data you have collected and sorted through, you will need to do a few other parts. These are all meant to help the person who needs this information; they can read through it and see what is inside and what they can do with all of that data. It is how they can use the information to see what is going on, the complex relationships that are there, and so much more.

This means that we need to spend our time with some write-ups of the data, graphs, charts, and other ways to represent and show the data to those who need it the most. This will form one of the final steps that come with data analysis. These methods are designed to distill and refine the data so that the readers are then able to glean some interesting information from it without having to go back through the raw data and figure out what is there all on their own.

Summarizing the data in these steps will be critical, and it needs to be done in a good and steady manner. Doing this will be critical to supporting some of the arguments made with that data, as is presenting the data clearly and understandably. During this phase, we have to remember that it is not always possible that the person who needs that summary and who will use it to make some important decisions for the business will be a data scientist. They need it all written out in simple and easy to understand this information. This is why the data has to be written out in a manner that is easy to understand and read through.

Often this is going to be done with some sort of data visualization. There are many visuals choices that we can use, and work with some kind of graph or

chart is a good option. Laboring with the best method for your needs and the data that we are using will be the best way to determine the visual that will be the best for you.

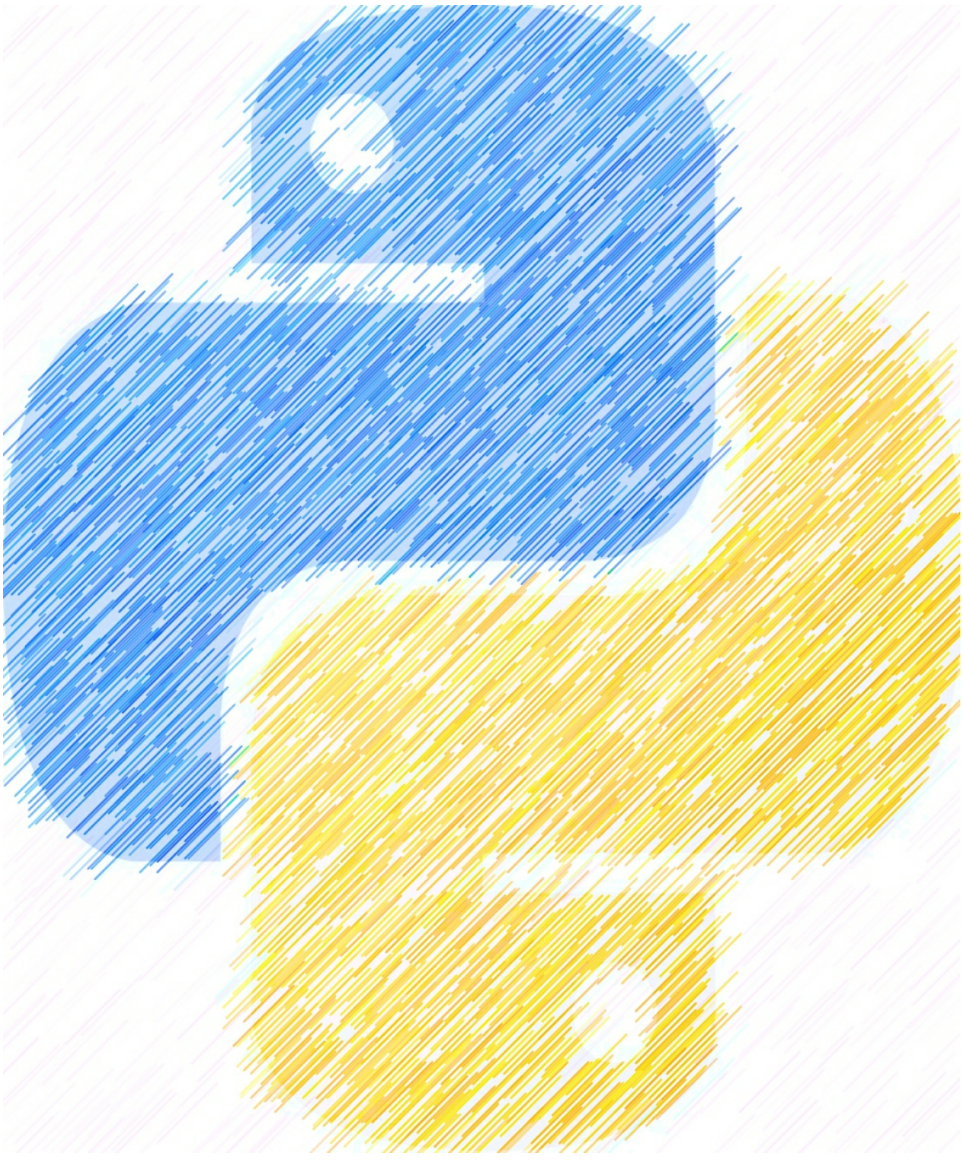
Reading through information in a more graphical format is going to be easier to work with than just reading through the data and hoping it to work the best way possible. You could have it all in a written form if you would like, but this will not be as easy to read through nor as efficient. To see some of those complex relationships quickly and efficiently, working with a visual will be one of the best options to choose.

Even though we need to spend some time working with a visual of the data to make it easier to work with and understand, it is fine to add some of the raw data as the appendix, rather than just throwing it out. This allows the person who will work with that data regularly a chance to check your resources and your specific numbers. It can help to bolster some of the results that you are getting overall.

If you are the one who is getting the results of the data analysis, make sure that when you get the conclusions and the summarized data from your data scientist that you go through and view them more critically. You should take the time to ask where the data comes from; it is going to be important, and you should also take some time to ask about the method of sampling used for all of this, including when the data were collected. Knowing the size of the sample is important, as well.

This will allow you to learn more about the data you have and then allow you to figure out if you can use the data or some bias that comes with it along the way. If the source of the data, or at least one of the sources, seems to have some kind of conflict that you are worried about, this will pull your results into question, and you at least need to look it over.

Chapter 2. Why Python for Data Analysis



How Python Can Help With Data Analysis

Now that we have had some time to discuss some of the benefits that come with the Python language and some of the parts that make up this coding language, it is now time for us to learn a few of the reasons why Python is the coding language to help out with all of the complexities and programs that we want to do with data science.

Looking back, we can see that Python has been pretty famous with data scientists for a long time. Although this language was not built just specifically to help out with data science, it is a language accepted readily and implemented by data scientists for much of the work they try to accomplish. Of course, we can imagine some of the apparent reasons why Python is one of the most popular programming languages and why it works so well with data science, but some of the best benefits of using Python to help out with your data science model or project include:

Python is as simple as it gets. One of the best parts about learning how to work with the Python coding language is that even as someone who is entirely new to programming and has never done any work in this past, you can grasp the basics of it pretty quickly. In particular, this language had two main ideas in mind when it was first started, and these include readability and simplicity.

These features are pretty unique when we talk about coding languages. They are often only going to apply to an object-oriented coding language and one that has a tremendous amount of potential for problem-solving.

All of this means that if you are a beginner to working with data science and working on the Python language, then adding these two together could be the key to getting started. They are both going to seem like simple processes when they work together, and yet you can get a ton done in a short amount of time. Even if you are more experienced with coding, you will find that Python data science will add a lot of depth to your resume and help you get those projects done.

The next benefit is that Python is fast and attractive. Apart from being as simple as possible, the code that we can write with Python will be leaner and much better looking than others. For example, the Python code takes up one-

third of the volume that we see with code in Java, and one-fifth of the volume of code in C++, just to do the same task.

The use of the common expressions in code writing, rather than going with variable declarations and space in place of ugly brackets, can also help Python's code to look better. In addition to having the code look more attractive, it can help take some of the tediousness that comes in when learning a new coding language. This coding language can save a lot of time and tax the brain of the data scientist a lot less, making working on some of the more complex tasks, like those of Data Analysis, much easier to handle overall.

The Python Data Analysis library, known as Pandas, is one of the best for helping us handle all of the parts of our Data Analysis and the whole process of data science. Pandas can grab onto many data without worrying about lagging and other issues in the process. This is great news for the data scientist because it helps them filter, sort, and quickly display their data.

Next on the list is that the Python library is quickly growing in demand. While the demand for professionals in the world of IT has seen a decline recently, at least compared to what it was in the past, the demand for programmers who can work with Python is steadily on the rise. This is good news for those who still want to work in this field and are looking for their niche or way to stand out when getting a new job.

Since Python has so many great benefits and has been able to prove itself a great language for many things, including programs for data analytics and machine learning algorithms, many companies centered on data will be into those with Python skills. If you already have a strong in Python, you can ride the market out there right now.

Finally, we come back to the idea of the vibrant community that is available with the Python language. There are times when you will work on a project, and things are just not working the way you had thought they would or the way you had planned. Getting frustrated is one option, but it is not going to help you to find the solution.

The good news with this is that you will be able to use the vibrant community and all of the programmers in this community to provide you with a helping

hand when you get stuck. The community that is around Python has grown so big, and it includes members who are passionate and very active in these communities

How Python Fits Into Data Analysis

The next thing on our list that we need to focus on is how we can work with Python to complete the data analysis that we would like. Many different parts come with our data analysis and having it all come together, and it will take us some time and some good planning in the process. At one point, though, we will need to go through and make sure that we are working with a programming language, versatile and strong, and one that will help us run our algorithms as we go.

Our algorithms are very important to how well the data analysis will work. These are the pars that will take ahold of our data and look through it all, sorting it through and telling us the insights or the patterns inside it. But to get these to work well and make sure that we are not going to end up with a big mess in the end and inaccurate results, we need to make sure that we choose a good and strong language to get it done.

There are many different coding languages that we can work with, and each one is going to bring about its positives and negatives that we need to deal with. If you hear about the idea of coding and learning how to do a programming language, and it makes you nervous and anxious, have no fear. There are many different languages that we can focus on to help us handle our algorithms and get the best results when we want to work with our data analysis.

The number one language that will work for data analysis, and the machine learning that we need to accomplish to handle these algorithms, is Python. As we will explore in this chapter, many benefits come with using Python, whether you want to learn the basics of coding or you are interested in handling something as complicated as data analysis. Let’s dive in and see what some of these benefits are all about.

There are many options in coding languages out there, but many of them will be kind of difficult to learn. They are often reserved for some of the more complicated types of coding you want to use, and you can build them later.

But if you are a complete beginner in coding, then Python will be the best option for you.

Python has a large library that makes learning the codes easier. You will be amazed at how much power will be found when you work with Python and how many options and functions are found in this language. Whether you are a beginner or looking to add a few other parts and coding languages to your skill set, you will find that the traditional Python library will have all of the parts you need to be more successful.

There are many extensions and other libraries that work with Python that are specifically designed to enhance its capabilities and make it work better for a good data analysis. Even though the traditional library that comes with Python will include a lot of the power and more that you want with coding, other extensions make sure you can complete some of the processes you want with data science, data analysis, and even machine learning. More than any other language, Python has many of these options, which can make it so much easier overall to get your work done.

There is a lot of power that you can enjoy when it is time to work on Python. Even though we have spent some time talking about how easy the Python language will be to learn, we have to remember that ease of use does not mean that you are missing out on power. The good news is that Python will come with a lot of power, and you will be able to use it to handle almost any project that you would like along the way.

The Python community is going to be large, allowing even a beginner to get some of the assistance that they need along the way. It may not seem like a big deal, but when you are working on learning how to work with a new language, it is going to prove to be invaluable. Any time you need to learn something new, you have a new question, or you get stuck, and you cannot figure out how to get things fixed and to work again, that community is going to be the answer you need.

The community is going to include programmers from all around the world. They will often have a lot of different experience levels when it comes to how much they know how to do with coding. As a beginner, you can easily join and be included. And many programmers who are more advanced are willing to share some of their time and knowledge with you. This helps to

facilitate some of the work you want to accomplish and make it easier to learn something new.

Chapter 3. The Steps of Data Analysis

With some of the ideas of a Data Analysis defined above to show us why this is so important, it is time for us to look at some of the steps that are so important to this process. When we know a bit more about some of the steps of Data Analysis and what we can do with it, we are going to find why we should use this method of learning from Big Data and then ensuring that your business will be able to use this information to get further ahead of the competition.

For most businesses, there isn't going to be any problem with a lack of information. These businesses will suffer from having too much information to handle, and they are not sure what they are supposed to do with it. This over-amount of data will make it harder to come up with a clear decision based on the data, and that can be a problem as well. With so much data to go and sort through, we need to get something more from the data.

This means that we need to know that the data we have is right for the questions we want to answer. We need to know how we can draw some accurate conclusions from the data we are working with. Besides, we need data that will be able to take on and inform our decision-making process.

In all, we need to make sure that we have the best kind of Data Analysis set up and ready to go. With the right process and tools set up for our Data Analysis, something that may have seemed like too much initially and like an overwhelming amount of stuff to go through will become a simple process that is clear and easy as possible.

To help us get all of this done, we need to go through some of the basic steps needed to use data to make better decisions overall. There are many ways to divide this all up and make it work better for our needs, but we are going to divide this up into five steps that we can use and rely on to see some of the best results overall. Some of the steps that we can use to help make our Data Analysis more productive for better decision making in the company include:

- Defining your question

- Setting up clear measurement
- Collecting the data
- Analyzing the data
- Interpreting the results

Defining Your Question

The first step that we need to undertake when working on Data Analysis is defining the main question that we would like to handle. You should not just randomly work with the data on hand and hope that it shows you something because it will get you lost and confused in the process. You need to have a clear picture of where you want to go and what you would like to learn from data, and then work from there.

In your Data Analysis, you need to start with the right questions. Questions are important, but we need to make sure that they are concise, measurable, and clear. Design the questions to either qualify or disqualify some of the potential solutions you are looking for on a specific problem or opportunity.

For example, you may want to start with a problem that you can clearly define. Maybe you are a government contractor, and you find that your costs are rising quite a bit. Because of this, you are no longer able to submit a competitive contract for some of the work that you are doing. You will want to go through with this and figure out how to deal with the business problem.

Setting up Clear Measurements

The next thing that we need to be able to do here is to set up some clear priorities on your measurements; this is one that we will be able to break down into two subsets to help us out. The first part is that we need to decide what we want to measure, and then the second thing we need to decide is how to measure it.

Then, we are going to get to the decision that we want to measure. When doing this, we need to make sure that we consider any of the objections, the reasonable ones, of stakeholders and others who are working with the company. They may be worried about what would happen if you reduced the staff, and then there was a big surge in demand shortly afterward, and you

could not hire more people in the right amount of time.

Once we are done with that first step, it is time for us to make some decisions on how to measure. Thinking about how we can measure the data that we have will be just as important here, especially before we go through the phase of collecting data because the measuring process will either back up our analysis or discredits it later on. There are frequently questions of different questions that you are going to ask in this stage, but some of the most important ones to consider will include:

1. What is the time frame that we are looking at?
2. What is the unit of measure that we are relying on?
3. What factors are important to consider in all of this?

Collecting the Data

After we have had some time to go through and define our big problem and then work on the measurements that we are going to use, it is time for us to move on to collecting the data. With the question defined and the measurement priorities set, it is now time for us to go through and collect the data. As we organize and collect the data, there are going to be many important points that we must keep in mind, which will include:

1. Before you collect some new data, you need to determine what we need to work with. We can look through some of the existing databases and some of the existing sources that we have on hand. You need to go through and collect some data first because it is simple and easier and can save a lot of money. We can move out to some other sources later, as well, if we need more information.
2. During this process, we also need to determine what naming system and file storing system we would like to use; this is going to make it easier for your team members to collaborate. This process will save some time and prevent members of your team from wasting time and money by collecting the same kind of information more than once.
3. If you would like to gather up data through interviews and observations, you need to develop an interview template ahead of time. This will ensure that we can save some time and that some continuity goes on in this process.
4. Finally, we need to keep the collected data that we have as organized as possible. We can work with a log with the collection dates and add in the notes about sources as you would like. This should also include some data normalization that you may have performed as well. This is going to be important because it will validate the conclusions that you make down the road.

As you go through this process, we need to make sure that we are taking care of some of the data we are working with; this means that we need to get it all organized, the values handled, and the duplicates we took care of before you try to do some of the analysis we want to do.

Since you are getting the data from different sources and working with data that may be incomplete and not perfect along the way, we need to be careful here. It is hard to know whether the data will be perfect or that you can use it the way you want. Besides, if the information is in the wrong format or brings some errors or missing values, it will be hard to get the algorithm to work.

The first thing that we need to do with this is to ensure that the data is in the same format. Usually, the best way to handle this for us is to go through and put all of the information in a standardized database that we can look at. We can use this in our storage service and make sure that all of the data we bring is put through that database and ready to go.

From there, we are tenable to focus on dealing with some of the errors found in that data. We want to make sure that the outliers, the missing values, and the duplicates are gone. For the most part, the outliers are things that you will need to ignore and get rid of. If there are a number of these, and they all end up in the same spot in the process, you should look at this to see what is going on and if this is new information that you should pay attention to. However, you will find that they are not worth your time, and most of these should just be ignored.

From there, we are going to look at the missing values. When you get information from the real world, there will be times when there is a missing value in one of the parts you are working with. You can either delete these if there are just a few or go through and replace these missing values with the mean of the other values in this column or row. It is up to you what you would like to do with these missing values to ensure that your information is as accurate as possible.

Finally, we need to deal with some of the duplicate values present in the data set. If any duplicate values show up in some of the data, we will find that this will skew a lot of the numbers that we have and the results that we will get. This is why we need to take care of them, so we can see the true values that

are inside of them.

You can go through here and figure out how much of the duplicate you would like to keep and how much you would like to get rid of during your time. Usually, it is best to limit it as much as possible. Sometimes, this keeps duplicates down to just two, and sometimes, it means only making sure that all of the entries are only in there once.

Chapter 4. Libraries

Many developers nowadays prefer the usage of Python in their data analysis. Python is not only applied in data analysis but also in statistical techniques. Scientists, especially the ones dealing with data, also prefer using Python in data integration. That's the integration of Webb apps and other environment productions.

The features of Python have helped scientists to use it in Machine Learning. Examples of these qualities include consistent syntax, being flexible and even having a shorter time in development. It also can develop sophisticated models and has engines that could help in predictions.

The following are examples of essential libraries being used in our present.

Scikit – Learn

Scikit learns it is one of the best and a modern library in Machine Learning. It has the ability to supporting learning algorithms, especially unsupervised and supervised ones.

Examples of Scikit learn include the following.

- K-means
- Decision trees
- Linear and logistic regression
- Clustering

This kind of library has major components from NumPy and SciPy. Scikit learns they can add algorithm sets useful in Machine Learning and tasks related to data mining. That is, it helps in classification, clustering, and even regression analysis. There are also other tasks that this library can efficiently deliver. A good example includes ensemble methods, feature selection, and, more so, data transformation. It is good to understand that the pioneers or experts can easily apply this if they can be able to implement the complex

and sophisticated parts of the algorithms.

TensorFlow

It is a form of algorithm which involves deep learning. They are not always necessary, but one good thing about these algorithms is their ability to correct results when done right. It will also enable you to run your data on a CPU or GPU. That's, you can write data in the Python program, compile it, then run it on your central processing unit. Therefore, this gives you an easy time in performing your analysis. Again, there is no need for having these pieces of information written in C++ or instead of other levels such as CUDA.

TensorFlow uses nodes, especially the multi-layered ones. The nodes perform several tasks within the system, including employing networks such as artificial neutral, training, and even set up a high volume of datasets. Several search engines such as Google depend on this type of library. One main application of this is the identification of objects. Again, it helps in different Apps that deal with voice recognition.

Theano

Theano also forms a significant part of the Python library. Its vital tasks here are to help with anything related to numerical computation. We can also relate it to NumPy. It plays other roles, such as:

- Definition of mathematical expressions
- Assists in the optimization of mathematical calculation
- It promotes the evaluation of expressions related to numerical analysis.

The main objective of Theano is to give out efficient results. It is a faster Python library as it can perform calculations of intensive data up to 100 times. Therefore, it is good to note that Theano works best with GPU compared to the CPU of a computer. In most industries, the CEO and other personnel use Theano for deep learning. Also, they use it for computing complex and sophisticated tasks. All these became possible due to its processing speed. Due to the expansion of industries with a high demand for data computation techniques, many people opt for the latest version of this library. Remember, the latest one came to the limelight some years back. The

new version of Theano, that's version 1.0.0, had several improvements, interface changes, and composed of new features.

Pandas

Pandas is a very popular library and helps in the provision of data structures that are of high level and quality. The data provided here is simple and easy to use. Again, it's intuitive. It is composed of various sophisticated inbuilt methods that can perform tasks such as grouping and timing analysis. Another function is that it helps in a combination of data and also offering filtering options. Pandas can collect data from other sources such as Excel, CSV, and even SQL databases. It also can manipulate the collected data to undertake its operational roles within the industries. The Pandas library consists of two structures that enable it to perform its functions correctly. That is the Series, which has only one dimension and data frames that are characterized by being two-dimensional. The Pandas library has been regarded as the strongest and powerful Python library for the time being. Its main function is to help in data manipulation. Also, it has the power to export or import a wide range of data. It is applicable in various sectors, such as in the field of Data Science.

Pandas is effective in the following areas:

- Splitting of data
- Merging of two or more types of data
- Data aggregation
- Selecting or subsetting data
- Data reshaping

Diagrammatic Explanations

Series Dimensional

A	7

B	8
C	9
D	3
E	6
F	9

Data Frames Dimensional

	A	B	C	D
*0	0	0	0	0
*1	7	8	9	3
*2	14	16	18	6
*3	21	24	27	9
*4	28	32	36	12
*5	35	40	45	15

Applications of pandas in a real-life situation will enable you to perform the following:

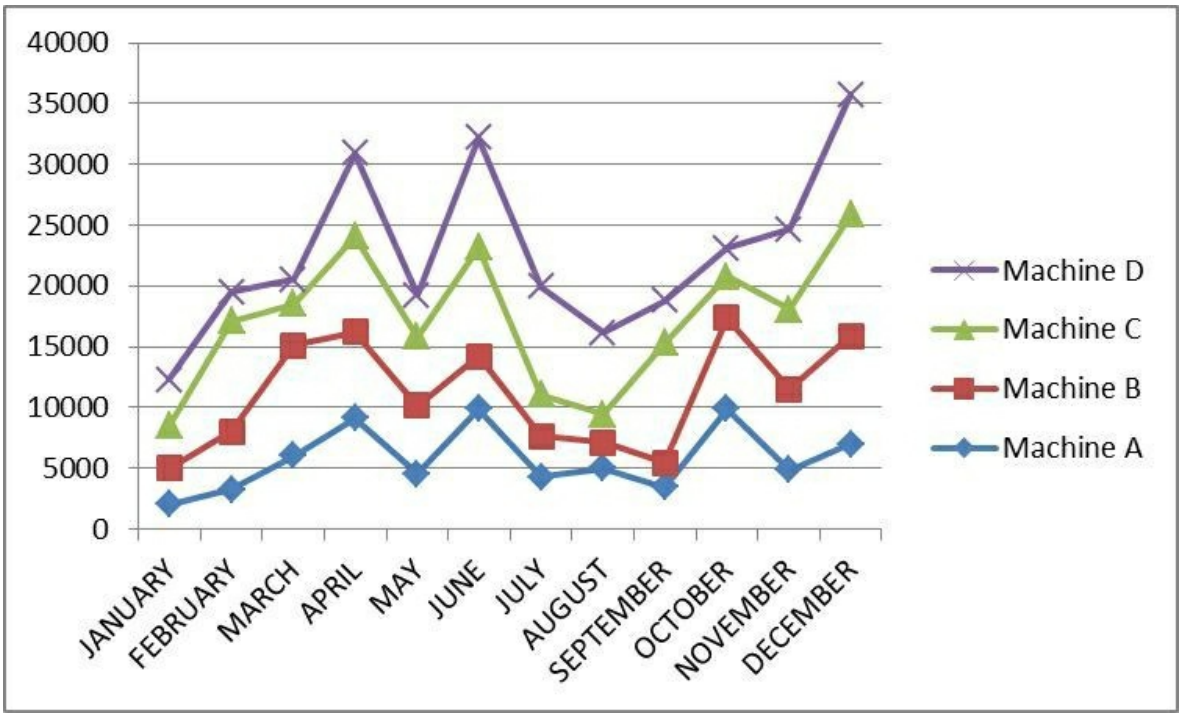
- You can quickly delete some columns or even add some texts found within the Dataframe
- It will help you in data conversion
- Pandas can reassure you of getting the misplaced or missing data
- It has a powerful ability, especially in the grouping of other programs according to their functionality.

Seaborn

Seaborn is also among the popular libraries within the Python category. Its main objective here is to help in visualization. It is important to note that this library borrows its foundation from Matplotlib. Due to its higher level, it is capable of various plot generation such as the production of heat maps, the

processing of violin plots, and the generation of time series plots.

Diagrammatic Illustrations



The above line graph clearly shows the performance of different machines the company is using. Following the diagram above, you can deduce and conclude which machines the company can keep using to get the maximum yield. On most occasions, this evaluation method will enable you to predict the exact abilities of your different inputs with the help of the Seaborn library. Again, this information can help for future reference in the case of purchasing more machines. Seaborn library also has the power to detect the performance of other variable inputs within the company. For example, the number of workers within the company can be easily identified with their corresponding working rate.

NumPy

This is a very widely used Python library. Its features enable it to perform multidimensional array processing. Also, it helps in the matrix processing.

However, these are only possible with the help of an extensive collection of mathematical functions. It is important to note that this Python library is highly useful in solving the most significant computations within the scientific sector. Again, NumPy is also applicable in linear algebra, derivation of random number abilities used within industries, and Fourier transformation. NumPy is also used by other high-end Python libraries such as TensorFlow for Tensors manipulation. In short, NumPy is mainly for calculations and data storage. You can also export or load data to Python since it has those features that enable it to perform these functions. It is also good to note that this Python library is also known as numerical Python.

SciPy

It is among the most popular libraries used in our industries today. It boasts of comprising of different modules that are applicable in the optimization sector of data analysis. It also plays a significant role in integration, linear algebra, and other forms of mathematical statistics.

 In many cases, it plays a vital role in image manipulation. Manipulation of the image is a process that is widely applicable in day to day activities; cases of Photoshop and much more are examples of SciPy. Again, many organizations prefer SciPy in their image manipulation, especially the pictures used for presentation. For

instance, wildlife society can come up with a cat's description and then manipulate it using different colors to suit their project. Below is an example that can help you understand this more straightforwardly.

The picture has been manipulated:

The original input image was a cat that the wildlife society took. After manipulation and resizing the image according to our preferences, we get a tinted image of a cat.

Koras

This is also part and parcel of the Python library, especially within Machine Learning. It belongs to the group of networks with high level neural. It is significant to note that Koras can work over other libraries, especially TensorFlow and even Theano. Also, it can operate nonstop without mechanical failure. In addition to this, it seems to work better on both the GPU and CPU. For most beginners in Python programming, Koras offers a secure pathway towards their ultimate understanding. They will be in a position to design the network and even to build it. Its ability to prototype faster and more quickly makes it the best Python library among the learners.

PyTorch

This is another accessible but open-source kind of Python library. As a result of its name, it boasts of having extensive choices when it comes to tools. It is also applicable in areas where we have computer vision. Computer vision and visual display play an essential role in several types of research. Again, it aids in the processing of Natural Language. More so, PyTorch can undertake some technical tasks that are for developers. That's enormous calculations and data analysis using computations. It can also help in graph creation, which is mainly used for computational purposes. Since it is an open-source Python library, it can work or perform tasks on other libraries such as Tensors. In combination with Tensors GPU, its acceleration will increase.

Scrapy

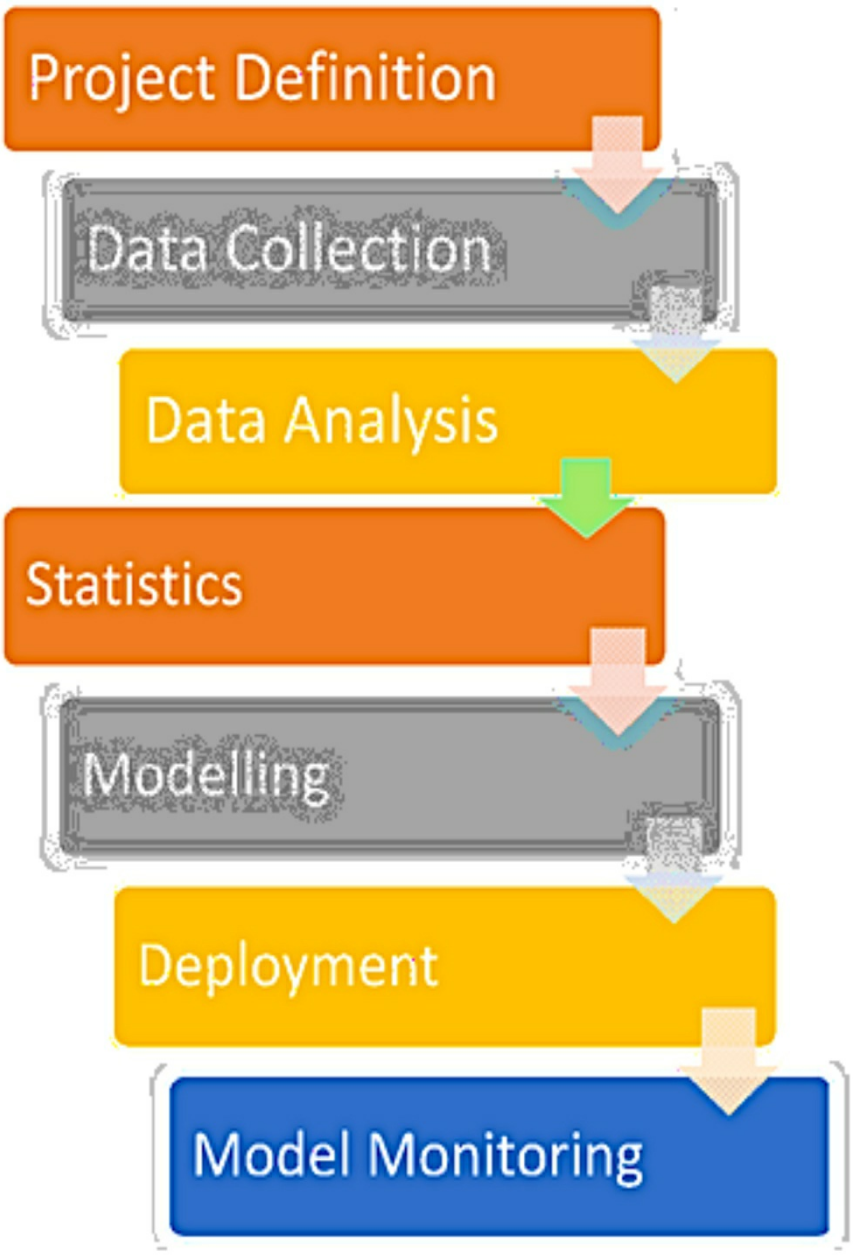
Scrapy is another library used for creating crawling programs. That's spider

bots and much more. The spider bots frequently help in data retrieval purposes and applicable in the formation of URLs used on the web. From the beginning, it was to assist in data scrapping. However, this has undergone several evolutions and led to the expansions of its general purpose. Therefore, the scrappy library's main task in our present-day is to act as crawlers for general use. The library led to the promotion of general usage, application of universal codes, and so on.

Statsmodels

Statsmodels is a library with the aim of data exploration using several statistical computations and data assertions. It has many features, such as result statistics and even characteristic features. It can undertake this role with various models such as linear regression, multiple estimators, and analysis involving time series, and even using more linear models. Also, other models, such as discrete choice, are applicable here.

Chapter 5. Predictive Analysis



What a Predictive Analysis Is

One topic that we need to explore a bit while we are here, before diving into how deep learning can help us out with it, is what predictive analytics is all about. Predictive analytics, to keep it simple, will be the use of techniques from machine learning, data, and statistical algorithms to help identify the likelihood of future outcomes based on more historical data. The goal is to go beyond what we know happened in the past to provide the best predictions and guess what will happen in the future.

Though the idea of predictive analytics is something that has been around for a long time, it is a technology that is starting to garner more attention than ever. Throughout a variety of industries, many companies are turning to predictive analytics to increase their bottom line and add a competitive advantage to everyone who uses it. Some of the reasons why this predictive analysis is gaining so much popularity now include:

1. The available data can help with this. The growing volume, as well as the types of data, are a good place to start. And there is more interest from many companies in using data to help produce some great insights in the process.
2. The computers and systems needed to complete this predictive analysis are cheaper and faster than ever before.
3. The software to finish the predictive analysis is easier to use.
4. There is a lot of competition out there for companies to work again. Thanks to these tougher conditions in the economy, and with the competition, businesses need to find their way to differentiate and become better than the competition. Predictive analysis can help them to do this.

Predictive analytics, with all of the interactive software that is easier than ever to use, has grown so much. It is no longer just the domain of those who study math and statistics. Business experts and even business analysts can use this kind of technology as well.

Keep in mind here that this is going to be a bit different compared to the descriptive models that can help us understand what happened in the past, or some of the diagnostic models that we can use that help us understand some key relationships and determine why we say a certain situation happens in the past. Entire books will be devoted to the various techniques and methods that are more analytical than others. And there are even complete college

curriculums that will dive into this subject as well, but we can take a look at some of the basics that come with this process and how we can use this for our needs.

There are two types of predictive models that we can take a look at first. These are going to include the classification models and regression models. To start with are the classification models that work to predict the membership of a class. For example, you may work on a project to figure out whether an employee is likely to leave the company, whether a person will respond to solicitation from your business, or whether the person has good or bad credit with them before you loan out some money.

We will stick with binary options for this kind of model, which means the results have to come in at a 0 or a 1. The model results will have these numbers, and 1 tells us that the event you are targeting is likely to happen. This can be a great way to make sure that we see whether something is likely to happen or not.

Then we have the regression models. These are going to be responsible for predicting a number for us. A good example of this would be predicting how much revenue a customer will generate over the next year, or how many months we have before a piece of our equipment will start to fail on a machine so you can replace it.

There are a lot of different techniques that we can work with when it comes to predictive modeling. The three most common types of techniques that fall into this category of predictive modeling will include regression, decision trees, and neural networks. Let’s take a look at some of these to see how these can all work together.

First on the list is a decision tree. It is an example of classification models that we can take a look at. This one will partition the data we want to work with and put it into subsets based on categories of the variables that we use as input. A decision tree will look like a tree that has each branch representing one of the choices that we can make. When we set this up properly, it can help us see how we want to make each choice compared to the alternatives. Each leaf out of this decision tree will represent a decision or a classification of the problem.

This model is helpful to work with because it looks at the data presented to it and then tries to find the single variable that can split up the data. We want to make sure that the data is split up into logical groups that are the most different.

The decision tree is going to be popular because they are easy to interpret and understand. They are also going to do well when it is time to handle missing values and are useful for the preliminary selection of your data. If you are working with a set of data that is missing many values or you would like a quick and easy answer that you can interpret in no time, then a decision tree is a good one to work with.

Then we need to move on to the regression. We are going to take a look at logistic and linear regression. The regression will be one of the most popular models to work with. The regression analysis will estimate the relationship that is present among the variables. It is also intended for continuous data that can be assumed to follow a normal distribution, it finds any of the big patterns that are found in sets of data, and it is often going to be used to determine some of the specific factors that answer our business questions, such as the price that can influence the movement of an asset.

As we work through the regression analysis, we want to make sure that we can predict a number called the response, or the Y variable. With some linear regressions, we will have one independent variable that can help explain or else predict the outcome of Y. The multiple regression will then work with two or more independent variables to help us predict what the outcome will be.

Then we can move on to the logistic regression. With this one, we will see that it is the unknown variable of a discrete variable that is predicted based on the known value of some of the variables. The response variable will be more categorical, which means that it can assume only a limited number of values compared to the others.

And finally, we have the binary logistic regression. This one is going to be a response variable that has only two values that go with it. All of the results that happen will come out as either 0 or as a 1. If we see 0, this means that the expected result is not going to happen. And if it shows up as a 1, then this means that our expected result will happen.

And then, we can end with the neural networks, as we talked about before. This will be a more sophisticated technique that we can work on and can model complex relationships. These are popular for many reasons, but one of the biggest is that neural networks are very flexible and powerful.

The power that we can see with the neural network will come with the ability that these have with handling nonlinear relationships in data, which is going to become more and more common as we work to collect some more data. Often, a data scientist will choose to work with the neural network to help confirm the findings that come with the other techniques you used, including the decision trees and regression.

Another option that we have to look at is artificial neural networks. These were originally developed by researchers trying to mimic what we can find in a human brain on a machine. And when they were successful, we get many modern pieces of technology that we like to use today.

Predictive analysis is going to do a lot of great things for your business. It can ensure that you will handle some of the more complex problems that your business faces and make it easier to know what is likely to happen in the future. Basing your big business decisions on data and the insights found with predictive analysis can make it easier to beat out the competition and get ahead.

Chapter 6. Combining Libraries

The PyTorch Library

The next library that we need to look at is known as PyTorch. This is a Python-based package that works for scientific computing.

It is going to rely on the power that it can receive from graphics processing units. This library will also be one of the most common and the preferred deep learning platforms for research to provide us with maximum flexibility and a lot of speed in the process.

There are plenty of benefits that come with this library. It is known for providing two of the most high-level features out of all the other deep learning libraries. These will include tensor computation to support a strong GPU acceleration and build up the deep neural networks on a tape-based autograd system. Through Python, many different libraries can help us work with much artificial intelligence and deep learning projects that we want to work with. The PyTorch library is one of these. One of the key reasons that this library is so successful is because it is completely Pythonic and one that can take some of the models that you want to build with a neural network almost effortlessly. This is a newer deep learning library to work with, but there is a lot of momentum in this field.

The Beginnings of PyTorch

Even though it was just released in January 2016, it has become one of the go-to libraries that data scientists like to work with, mainly because it can make it easy to build complex neural networks. This is perfect for countless beginners who haven’t been able to work with these neural networks at all in the past. They can work with PyTorch and make their network in no time, even with a limited amount of coding experience.

This Python library’s creators envisioned that this library would be imperative when they wanted to run large numerical computations as quickly

as possible. This is one of the ideal methodologies that perfectly fits with the programming style that we see with Python. This library, along with the Python library, allowed debuggers of neural networks, machine learning developers, and deep learning scientists to run and test parts of their code in real-time. This is great news because it means that these professionals no longer have to wait for the entire code to complete and execute before they can check out whether this code works or if they need to fix certain parts.

In addition to some of the functions that come with the PyTorch library, remember that you can extend out some of this library's functions by adding in other Python packages. Python packages like Cython, SciPy, and NumPy all work well with PyTorch as well.

Even with these benefits, we still may have some questions about why the PyTorch library is so special and why we may want to use this when it is time to build up the needed models for deep learning. The answer to this is simple: PyTorch is seen as a dynamic library.

This means that the library is flexible, and you can use it with any requirements and changes that you would like. It is so good at doing this job that developers are using it in artificial intelligence, and by students and researchers in many industries. In fact, in a Kaggle competition, this library was used by almost all of the individuals who finished in the top ten.

While there are multiplied benefits that can come with the PyTorch library, we need to start with some of the highlights of why professionals of all sorts love this language so much.

Reasons to Use PyTorch With the Data Analysis

Any individual working in information science, information investigation, human-made reasoning, or profound learning has likely invested some energy working with the TensorFlow library, which we have discussed in this manual. TensorFlow might be the most famous library from Google. Still, since the PyTorch system for profound learning, we can find that this library can take care of a couple of new issues regarding investigating work that these experts need to fix.

It is frequently accepted that PyTorch is currently the greatest contender out

there to TensorFlow with regards to taking care of information, and it is truly outstanding and most loved human-made brainpower and profound learning library with regards to the network of examination. There are numerous purposes behind this incident, and we will discuss a portion of these beneath.

To start with, we will see that the dynamic computational diagrams are mainstream among analysts. This library will dodge a portion of the static diagrams utilized in different structures from TensorFlow. This permits analysts and engineers to change how the organization is finally set.

Some of those receiving this library will like it because these charts are more intuitive to realize when we contrast them with what TensorFlow can do.

The following advantage is that this one accompanies an alternate sort of backend uphold. PyTorch will utilize an alternate backend dependent on what you are doing. The GPU, CPU, and other practical highlights will all accompany an alternate backend, instead of zeroing in on only one backend to deal with these. For instance, we will see the THC for our GPU and the TH for CPU.

Having the option to utilize separate backends can make it simpler to convey this library through an assortment of compelled frameworks. The basic style is another advantage of working with this sort of library. This implies that it is difficult to utilize when we work with this library and is extremely intuitive. When you execute a line of code, it will get similarly executed as you need, and you can work with the continuous following. This permits the developer to monitor how the models for neural organizations are doing. As a result of the incredible design and the lean and quick methodology, we have had the option to expand a portion of the available selections that we see with this library all through developers' networks. Another advantage that we will appreciate about working with PyTorch is that it is anything but difficult to broaden. This library, specifically, is incorporated to function admirably with the code for C++. It will share a touch of the backend with this language when we take a shot at our system for profound learning.

Pandas

Pandas are built on NumPy, and they are meant to be used together. This makes it extremely easy to extract arrays from the data frames. Once these

arrays are extracted, they can be turned into data frames themselves.

Matrix Operations

This includes matrix calculations, such as matrix-to-matrix multiplication. Let's create a two-dimensional array.

This is a two-dimensional array of numbers from 0 to 24. Next, we will declare a vector of coefficients and a column that will stack the vector and its reverse.

Slicing and Indexing

Indexing is great for viewing the nd-array by sending instructions to visualize the slice of columns and rows or the index.

This is one of the most important libraries that we can work with overall because it can handle pretty much all of the parts that come with data analysis. There isn't anything in data analysis that the Pandas library won't help us out with. Pandas are one of Python's packages that can provide us with numerous different tools to help in data analysis. The package will come with a lot of different data structures that can be used for the different tasks that we need to do to manipulate our data. It will also come with a lot of methods that we can invoke for the analysis, which is useful when we are ready to work on some of our machine learning and data science projects in this language.

As we can imagine already, there are several benefits that we can enjoy when we work with the Pandas library, especially when compared to some of the other options out there. First, it will present our data to handle all of our analysis through the different data structures, particularly through the DataFrame and the Series structures.

In addition to this, we will find that this is a package that can contain many different methods convenient for data filtering and more. The Pandas library will come with many utilities that we need to perform operations of Input and Output seamlessly. And no matter which format your data will come to us in, whether it is CSV, MS Excel, or TSV, the Pandas library can handle it for us.

Chapter 7. Machine Learning and Data Analysis

What Machine Learning Is

The first thing that we need to take a look at here is the basics of machine learning. This will be one of the techniques that we can use with data analytics that will help teach a computer how to learn and react on their own without the programmer's interaction. Many of the actions that we will train the system to do will be similar to actions that already come naturally to humans, such as learning from experience.

The algorithms that come with machine learning will be able to use computational methods to learn information right from the data without having to rely on an equation that is predetermined as its model. The algorithms will adaptively improve some of their performance as the number of samples we will use for learning will increase.

There are a lot of instances where we can use machine learning. With the rise in big data that is available for all industries to use, We will find that machine learning is going to become one of the great techniques that are used to solve a ton of problems in many areas, including the following:

1. **Computational finance:** This will include algorithmic trading, credit scoring, and fraud detection.
2. **Computer vision and other parts of image processing:** This can be used in some different parts like object detection, motion detection, and face recognition.
3. **Computational biology:** This will be used for a lot of different parts, including DNA sequencing, drug discovery, and tumor detection.
4. **Energy production:** This can help with a few different actions like load forecasting and help predict what the prices will be.
5. **Manufacturing, aerospace, and automotive options:** This will be a great technique to work with when it comes to helping with many parts, including predictive maintenance.

6. **Natural language processing:** This will be the way that we can use machine learning to help with voice recognition applications.

Machine learning and the algorithms that they control will work by finding some natural patterns in the data that you can use, including using it in a manner that will help us make some better predictions and decisions along the way. They will be used daily by businesses and a lot of different companies to make lots of critical decisions.

For example, medical facilities can use this to help them to help diagnose patients. And we will find that many media sites will rely on machine learning through the potential of millions of options to give recommendations to the users. Retailers can use this to gain some insight into the purchasing behavior of their customers along the way.

There are many reasons that your business can consider using machine learning. For example, it will be useful if you are working with a complex or one that is going to involve a larger amount of data and a ton of variables. Still, there isn't an equation or a formula out there right now to handle it. For example, some of the times when we want to work with machine learning include:

1. Equations and rules that are hand-written and too complex to work with. This could include some options like speech recognition and face recognition.
2. When you find that the rules are going to change all of the time, this could be seen in actions lie fraud detection from a large number of transactional records.
3. When you find that your data's nature is going to change constantly, and the program has to be able to adapt along the way. This could be seen when we predict the trends during shopping when doing energy demand forecasting and even automated trading, to name a few.

As you can see, there are many different things that we can do when it comes to machine learning, and pretty much any industry will be able to benefit from working with this for their own needs. Machine learning is more complex, but we can do it with Python for some amazing results in the process and ensure our data analysis will work the way that we want

Decision Trees and Random Forests

Decision trees are algorithms that try to classify the elements by identifying

questions concerning their attributes that will help decide which class is the correct to place them. Each node inside the tree is a question, with branches that lead to more questions about the articles and the leaves as the final classifications.

Use cases for decision trees can include the construction of knowledge management platforms for customer service, price predictions, and product planning.

An insurance agency could utilize a decision tree when it requires data about the sort of protection items and the tremendous changes dependent on possible hazard, says Ray Johnson, boss information researcher at business and innovation counseling firm SPR. Utilizing area information overlaid with climate-related misfortune information, you can make hazard classes dependent on claims made and cost sums. It would then assess new models of the fence against models to give a hazard class and a potential monetary effect, the official said.

Random Forests; a decision tree must be prepared to give precise outcomes. The rough timberland calculation takes many irregular choice trees that base their choices on various arrangements of attributes and permit them to cast a ballot in the most well-known request.

Random forests are simply flexible devices for discovering connections in data sets and quick to train, says Epstein. For example, unsolicited bulk mail has been a problem for a long time, not only for users but also for Internet service providers who manage the increased load on servers. In response to this problem, automated methods have been developed to filter spam from standard email, using random forests to quickly and accurately identify unwanted emails, the executive said.

Other uses of random forests include identifying a disease by analyzing the patient’s medical records, detecting bank fraud, predicting the volume of calls in the call centers, and predicting gains or losses through the Purchase of a particular stock.

SciKit-Learn

This is a fundamental tool used in data-mining and data analysis related tasks. This is an open-source tool and licensed under BSD. This tool can be accessed or reused in different contexts. SciKit has been developed on top of NumPy, Matplotlib, and SciPy. The tool is utilized for classification, regression, clustering, and managing spam, image recognition, stock pricing, drug response, customer segmentation, etc. The tool also permits model selection, dimensionality reduction, and pre-processing.

Linear Regression

The word “linearity” in algebra implies a linear connection between two or more variables. We get a conservative line if we draw this connection in a two-dimensional space (amongst two variables).

Linear regression completes the duty to foresee a dependent variable rate (y) built on a certain independent variable (x). So, this regression method finds out a linear connection between x (input) in addition to y (output). Thus, they term it Linear Regression. Suppose we plot the dependent and independent variables (y and x) on their axis. In that case, linear regression gives us a conventional line that fits the information plugs, as revealed in the picture below. We then recognize that the equation of a conventional line is essential.

The equation of the overhead line is:

$$Y= mx + b$$

Where b is the advert and m are the hills of the line. So, essentially, the linear regression algorithm gives us the most significant ideal rate for the advert and the hill (in two magnitudes). Although they and x variables produce the result, they are the data structures and cannot be altered. The figures that we can switch are the advert(b) and hill(m). There can be numerous conventional lines relying upon the figures of the advert and the hill figures. Essentially, the linear regression algorithm ensures it fits numerous lines on the data points and yields the line that results in the slightest mistake.

This similar idea can be stretched to cases where there are additional variables. This is termed numerous linear regressions. For example, think

about a situation where you must guess the house's price built upon its extent, the number of bedrooms, the regular income of the people in the area, the oldness of the house, and so on. In this situation, the dependent variable (target variable) is reliant on numerous independent variables. A regression model, including numerous variables, can be signified as:

$$y = b0 + m1b1 + m2b2 + m3b3 + ... \text{ mean}$$

This is the comparison of a hyperplane. Recall that a linear regression model in two magnitudes is a straight line; in three magnitudes, it is a plane, and in additional magnitudes, a hyperplane.

Support Vector Machines (SVM)

A managed algorithm used for machine learning which can mutually be employed for regression or classification challenges is Support Vector Machines. Nevertheless, it is typically employed in classification complications. In this algorithm, we design each data entry as a point in n-dimensional space (where n is many structures you have), with the rate of each feature being the rate of a coordinate. Then, we complete the classification by finding the hyper-plane that distinguishes the two classes very well.

K-means Clustering

The 2000 and 2004 Constitutional determinations in the United States were closed. The highest percentage received by any runner from a general ballot was 50.7%, and the lowest was 47.9%. If a proportion of the electorates were to have their sides swapped, the determination would have been dissimilar. There are small clusters of electorates who, when appropriately enticed, will change sides. These clusters may not be gigantic, but they might be big enough to change the result of the determination with such close competitions. By what means do you find these clusters of individuals? By what means do you request them with an inadequate budget? To do this, you can employ clustering.

Let us recognize how it is done.

- First, you gather data on individuals either with or without their permission: any kind of data that might give an approximate clue about what is vital to them and what will affect how they vote.
- Then you set this data into a clustering algorithm.
- Next, for each group (it would be very nifty to select the principal one first), you create a letter to appeal to these electorates.
- Lastly, you send the campaign and measure to see if it's employed.

Clustering is a category of unsupervised learning that routinely makes clusters of comparable groups. It is like an involuntary classification. You can cluster nearly everything, and the more comparable the objects are in the cluster, the enhanced the clusters are.

Chapter 8. Applications

Before we are finished with this manual, we need to take a gander at a portion of the applications that will assist us with taking advantage of information investigation. There are countless ways that this information investigation will be utilized, and when we can assemble it all, we will see some great outcomes all the while. Spots like the monetary world, security, promoting, publicizing, and medical services are largely going to profit by this information investigation, and as additional time goes on, all things considered, we will see a greater amount of these applications too. A portion of the manners in which we can work with information investigation and get the best outcomes from it include:

Security

A few urban communities worldwide are dealing with a prescient examination so they can foresee the zones of the town where there is bound to be a major flood for wrongdoing that is there. This is finished with the assistance of some information from an earlier time and even information on the zone's geology.

This is something that a couple of urban areas in America have had the option to utilize, including Chicago. Even though we can envision that it is difficult to utilize this to get each wrongdoing that is out there, the information that is accessible from utilizing this will make it simpler for cops to be available in the perfect regions at the perfect occasions to help lessen the rates of bad actions in a portion of those regions. Also, later on, you will find that when we use information investigation in this sort of way, in the huge urban communities, it has assisted with making these urban areas and these territories significantly more secure. The dangers would not need to put their lives at risk as much as in the past.

Transportation

The universe of transportation can work with information investigation, also. A couple of years back, when plans were being made at the London Olympics, there was a need to deal with more than 18 million excursions made by fans into the city of London. Also, it was something that we had the option to figure out well.

How was this commitment accomplished for these individuals? The train administrators and the TFL administrators worked with information investigation to ensure that every one of those excursions went as easily as could reasonably be expected. These gatherings had the option to experience and being informed from the occasions around that time and afterward utilized this as an approach to estimate the number of individuals who would head out to it. This arrangement went so well that the observers and the competitors could be moved to and from the correct spots conveniently the entire occasion.

Danger and Fraud Detection

This was one of the first employments of information investigation and was frequently utilized in the accounts. Numerous associations had a terrible involvement underwater, and they were prepared to roll out certain improvements to this. Since they had a hang on the information gathered each time the client came in for an advance, they could work with this cycle not to lose so much cash.

This permitted the banks and other monetary foundations to jump and conquer a portion of the information from the profiles they could use from those clients. The moment the bank or monetary foundation can dispose of the clients they work with, the costs that have arisen recently, and some of the other data that is significant for these agencies, they will decide on some better options about whom to give cash credit to, greatly diminishing their dangers. This encourages them to offer better rates to their clients.

Notwithstanding helping these monetary foundations ensure that they can distribute advances to clients bound to repay them, you will find that this can be utilized to help cut down on the dangers of extortion. This can cost the

bank billions of dollars a year and can be costly to work with. When the bank can utilize the entirety of the information that they have for finding exchanges that are false and making it simpler for their clients to keep cash in their record, they can ensure that they will not lose money in the process as well.

Coordination of Deliveries

There are no impediments with regards to what we can do with our information investigation, and we will find that it functions admirably with regards to coordination and conveyances. A few organizations focus on coordination, which will work with this information investigation, including UPS, FedEx, and DHL. They will utilize information to improve how effective their tasks are.

From the use of the information, it is feasible for these organizations that use it to locate the best and most productive courses to use when dispatching the things, which will guarantee that the things are transported on time, therefore substantially more. This causes the item to get something through in seconds and minimizes expenses to a base. Alongside this, the data that the organizations can assemble through their GPS can give them more open doors later on to utilize information science and information investigation.

Client Interactions

Numerous organizations will work with the use of information investigation to have better communications with their clients. Organizations can do a ton about their clients, frequently with some client overviews. For instance, numerous insurance agencies will utilize this by conveying client reviews after connecting with their controller. The insurance agency is then ready to use which of their administrations are acceptable, that the clients like, and which ones they might want to take a shot to see a few upgrades.

There are many socio-economic aspects that a company can work with. It is conceivable that these will require numerous assorted techniques for correspondence, including email, telephone, sites, and in-person communications. Taking a portion of the examination that they can get with their clients' socioeconomics and the input that comes in will guarantee that

these insurance agencies can offer the correct items to these clients. It depends 100% on the demonstrated bits of knowledge and client conduct also.

City Planning

One of the serious mix-ups made in numerous spots is that investigation, particularly the means that we are discussing in this manual, isn't something that is being utilized and thought about regarding city arranging. Web traffic and advertising are the things that are being used rather than the production of structures and spaces; this will cause huge numbers of issues that will come up when we talk about the control over our information because there are a few impacts overbuilding drafting and making new things on the road in the city.

Models that have been fabricated well will help amplify the openness of explicit administrations and territories while guaranteeing that there isn't the danger of over-burdening huge components of the city's framework simultaneously. This helps ensure a degree of effectiveness as everybody, however much as could reasonably be expected, can get what they need without doing a lot to the city and causing hurt like this.

We will typically observe structures not placed in the correct spots or organizations that are moved where they don't have a place. How frequently have you seen a structure that was on a detect that seemed as though it was reasonable and useful for the need, however, which had a ton of negative effect on different spots around it? This is because these potential issues were not a piece of thought during the arranging time frame. The uses of information research, and some demonstrations, make things simpler because we will realize what could happen if we put that building or something else in which it is recognized that there is a choice to be made.

Medical Care

The medical care industry has had the option to see numerous advantages from information investigation. However, there are numerous techniques; we will take a gander at one of the primary difficulties that emergency clinics will confront. Additionally, they need to adapt to cost pressures when they

need to treat; however, many patients could reasonably be expected while still getting excellent consideration from the doctors. This makes the specialists and other staff fall behind in a portion of their work every so often, and it is difficult to stay aware of the interest.

You will find that the information we can use here has risen a lot, and it permits the clinic to advance and afterward track the treatment of their patient. It is also a decent method to follow the patient stream and how the emergency clinic's diverse gear is being used. Indeed, this is incredible to the point that it is assessed that using this information investigation could give a 1 percent productivity pick up and could bring about more than $63 billion in overall medical care administrations. Consider what that could intend to you and everyone around you.

Specialists will work with information investigation to give them an approach to help their patients somewhat more. They can utilize this to analyze and comprehend what is new with their patients reasonably and more productively. This can permit specialists to furnish their clients with an excellent encounter and better consideration while guaranteeing that they can stay aware of all they require to do.

Travel

Information investigation and a portion of their applications are a decent method to enhance the purchasing experience for a traveler. This can be validated through an assortment of choices, including information investigation of portable sources, sites, or web-based media. The explanation behind this is that the longings and the inclinations of the client can be acquired from these sources, making organizations begin to sell out their items on account of the relationship of all the ongoing perusing on the site, including any of the money offers to help transformations of the buying habits. They can use the entirety of this to offer some modified bundles and offers. The uses of information investigation can likewise assist with conveying some customized travel suggestions. It regularly relies upon the result that the organization can get from their information via online media.

Computerized Advertising

Aside from simply using it to help some look through another, there is another area where we can witness an investigation of the information on a consistent basis, and that is computerized advertising. From a portion of the banners found in a few places to the advanced ads you might be used to finding in a portion of the largest and most urban communities, these will be controlled by the calculations of our information along the way.

Chapter 9. Data Visualization and Analysis With Python

Enormous Data

Huge information alludes to a torrential slide of organized and unstructured information perpetually flooding and from an assortment of unending information sources. These informational indexes are too huge to be broken down with conventional scientific apparatuses, and advancements have plenty of important bits of knowledge stowing away underneath.

The Versus of Big information

- **Volume:** To be named huge information, the given informational index's volume must be significantly bigger than conventional informational indexes. These informational indexes are fundamentally made out of unstructured information with restricted organized and semi-organized information. The unstructured information or the information with obscure worth can be gathered from input sources, for example, pages, search history, versatile applications, and web-based media stages. The organization's size and client base normally correspond to the volume of the information procured by the organization.
- **Speed:** The speed at which information can be assembled and followed up on the first to huge information speed. Organizations are progressively utilizing a mix of on-reason and cloud-based workers to speed up their information assortment. The present-day "Brilliant Products and Devices" require constant admittance to customer information to have the option to give them an additionally captivating and upgraded client experience.
- **Assortment:** Traditionally, an informational index would contain a larger part of organized information with a low volume of unstructured and semi-organized information. The enormous information approach has offered to ascend to new unstructured information types, such as video, text, and sound, that require complex instruments and innovations to clean and handle these information types to extricate important experiences.
- **Veracity:** Another "V" that must be considered for extensive information investigation is veracity. This alludes to the "dependability or the quality" of the information. For instance, web-based media stages like "Facebook" and "Twitter" with sites and posts containing a hashtag, abbreviations, and a wide range of composing mistakes can fundamentally diminish the unwavering quality and precision of the informational

indexes.

- **Worth:** Data has advanced as money with inherent worth. Much like conventional financial, monetary forms, a definitive estimation of the huge information corresponds to the understanding accumulated from it.

"The significance of enormous information doesn't spin around how much information you have, yet how you deal with it. You can take information from any source and investigate it to discover answers that empower 1) cost decreases, 2) time decreases, 3) new item advancement and upgraded contributions, and 4) intelligent dynamic."

SAS

The working of enormous information. There are three significant activities needed to pick up experiences from large information:

- **Coordination:** The conventional information reconciliation strategies, for example, ETL (Extract, Transform, and Load), are unequipped for gathering information from a wide assortment of outside sources and applications that are at the core of huge information. Progressed instruments and innovations are needed to break down huge informational indexes that are dramatically bigger than conventional informational collections. By incorporating enormous information from these sources, organizations can examine and remove significant knowledge to develop and keep up their organizations.
- **The executives:** Big information on the board can be characterized as "the association, organization, and administration of enormous volumes of both organized and unstructured information." Big information requires effective and modest capacity, which can be refined utilizing workers that are on-premises, cloud-based, or a blend of both. Organizations can flawlessly get to required information from anyplace over the world, and afterward, handling this is information utilizing required preparing motors dependent upon the situation. The objective is to ensure the nature of the information is significant and can be gotten effectively by the necessary devices and applications; large information is assembled from a wide range of data sources, including online media stages, web index history, and call logs. The enormous information typically contains huge arrangements of unstructured information and semi-organized information put away in an assortment of organizations. To have the option to measure and store this muddled information, organizations require all the more impressive and advanced than the dashboard programming beyond the usual social data sets and information distribution center stages.

New stages are accessible in the market that is equipped for joining enormous information with the conventional information distribution center frameworks in a "sensible information warehousing design." As a component of this effort, organizations need to decide which information should be secured for

administrative and consistency purposes, which information should be kept for future scientific purposes, and which information has no future and can be discarded. This cycle is designated "information characterization," which permits a fast and proficient investigation of a subset of information to be remembered for the organization's prompt dynamic cycle.

- **Investigation:** Once the enormous information has been gathered and effectively open, it may be dissected utilizing progressed logical apparatuses and innovations. This investigation will give important knowledge and significant data. Huge information can be investigated to make disclosures and create information models utilizing computerized reasoning and AI calculations.

Enormous Data Analytics

The particulars of enormous information and extensive information investigation are regularly utilized reciprocally because the intrinsic motivation behind huge information is dissected. "Enormous information investigation" can be characterized as a bunch of subjective and quantitative strategies that can be utilized to inspect a lot of unstructured, organized, and semi-organized information to find information examples and significant shrouded bits of knowledge. Large information investigation is the study of examining huge information to gather measurements, key execution markers, and Data drifts that can be effectively lost in the surge of crude information, purchase utilizing AI calculations, and insightful mechanized methods. The various advances engaged with "enormous information investigation" are:

- **Social occasion Data Requirements:** It is critical to comprehend what data or information should be accumulated to meet the business goal and objectives. Information association is likewise basic for proficient and exact information investigation. A portion of the classifications in which the information can be coordinated is the sexual orientation, age, socioeconomics, area, identity, and pay. A choice should likewise be made on the necessary information types (subjective and quantitative), and information esteems (mathematical or alphanumerical) to be utilized for the examination.
- **Get-together Data:** Raw information can be gathered from unique sources, for example, web-based media stages, PCs, cameras, other programming applications, organization sites, and even outsider information suppliers. The huge information investigation characteristically requires enormous volumes of information, mostly unstructured with a restricted measure of organized and semi-organized information.
- **Information association and order:** Depending on the organization's framework, data association should be possible on a specific Excel accounting page or by using and managing devices and applications suitable for handling factual information. The

information should be coordinated and organized according to the information needs gathered in synchrony with the research measure of the large information.

- **Cleaning the information:** It is critical to play out the extensive information investigation adequately and quickly to ensure the informational index is free of any excess and mistakes. Just a total informational index is satisfying the Data necessities more likely than not to continue to the last investigation step. Preprocessing of information is needed to ensure that the great main information is being dissected and that organization assets are being effectively utilized.
- **Examining the information:** Depending on the understanding that is required to be accomplished by the consummation of the investigation, any of the accompanying four unique kinds of huge information investigation approach can be embraced:

1. **Predictive investigation:** This sort of examination is done to produce figures and expectations for the organization's tentative arrangements. By the consummation of prescient investigation on the organization's huge information, the organization's future condition can be all the more correctly anticipated and got from the organization's present status. This investigation inspires the business chiefs to ensure that the organization's everyday activities follow its future vision. For instance, to send progressed scientific apparatuses and applications in an organization's business division, the initial step is to break down the main wellspring of information. Once accepts source examination has been finished, the sort and number of correspondence channels for the business group must be broken down. This is trailed by the utilization of AI calculations on client information to pick up knowledge into how the current client base is connecting with the organization's items or administrations. This prescient investigation will finish up with the arrangement of human-made reasoning based instruments to soar the organization's deals.
2. **Prescriptive examination:** The analysis is completed by principally zeroing in on the business rules and suggestions to create a particularly insightful way recommended by the business guidelines to support organization execution. This investigation aims to comprehend the complexities of different branches of the association and what measures should be taken by the organization to have the option to pick up bits of knowledge from its client information by utilizing a recommended insightful pathway. This permits the organization to grasp area particularity and compactness by sharply spotlighting its current and future enormous information investigation measures.

The big data analysis can be conducted using one or more of the tools listed below:

- **Hadoop:** Open source data framework.
- **Python:** Programming language widely used for machine learning.
- **SAS:** Advanced analytical tool used primarily for big data analysis.
- **Tableau:** Artificial intelligence-based tool used primarily for data visualization.
- **SQL:** the Programming language used to extract data from relational databases.
- **Splunk:** Analytical tool used to categorize machine-generated data

- **R-programming:** The Programming language used primarily for statistical computing.

Chapter 10. Data Science

Data Science and Its Significance

Data Science has come a long way from the past few years, and thus, it becomes an important factor in understanding the workings of multiple companies. Below are several explanations that prove data science will still be an integral part of the global market.

- The organizations would have the option to comprehend their customer in a more effective and high way with Data Science's assistance. Fulfilled clients structure each organization's establishment, and they assume a significant function in their victories or disappointments. Information Science permits organizations to draw in with clients in the development way and along these lines demonstrates the item's improved presentation and strength.
- Data Science empowers brands to convey ground-breaking and drawing in visuals. That is one reason it's renowned. When items and organizations utilize this information, they can impart their encounters to their crowds and make better relations with the thing.
- Perhaps one of Data Science's critical qualities is that its outcomes can be summed up to practically a wide range of ventures, such as travel, medical care, and schooling. The organizations can rapidly decide their issues with the assistance of Data Science and can likewise enough address them
- Currently, information science is available in practically all ventures, and these days, there is a colossal measure of information existing on the planet. Whenever utilized enough, it can prompt triumph or disappointment of any task. If information is utilized appropriately, it will be significant later on to accomplish the item's objectives.
- Big information is consistently on the ascent and developing. Huge information permits the undertaking to adequately address convoluted business, human resources, and capital administration issues and rapidly utilize various assets assembled regularly.
- Data science is picking up quick prevalence in each area and, in this way, assumes a significant part in each item's working and execution. In this way, the information researcher's job is improved as they will lead a basic capacity to oversee information and give answers for specific issues.
- Computer innovation has likewise influenced store areas. To get this present, we should take a model the more established individuals had a special connection with the neighborhood merchant. Likewise, the dealer had the option to meet the clients' necessities in a customized way. Be that as it may, presently, this consideration was lost because of chain stores' development and increment. Be that as it may, the vendors

can speak with their clients with information investigation assistance.

- Data Science assists organizations with building that client association. Organizations and their merchandise will have the option to have a superior and more profound comprehension of how customers can use their administrations with information science assistance.

Future of Information Technology

Like different territories are consistently advancing, information innovation's significance is progressively developing. Information science affected various fields. Its impact can be seen in numerous businesses, for example, retail, medical services, and training. New medicines and advances are, as a rule, constantly distinguished in the medical care area, and there is a requirement for quality patient care. The medical care industry can discover an answer with information science procedures that causes the patients to deal with. Schooling is another field where one can unmistakably observe the upside of information science. Presently the new developments like telephones and tablets have gotten a fundamental attribute of the instructive framework. Likewise, with the assistance of information science, the understudies make more prominent possibilities, which prompts improving their insight.

Information Structures

An information structure might be chosen in PC programming or intended to store information to work with various calculations. Every other information structure incorporates information, information connections, and capacities between the information that can be applied to the information and data.

Highlights of Information Structures

Here and there, information structures are classified by their attributes. Potential capacities are:

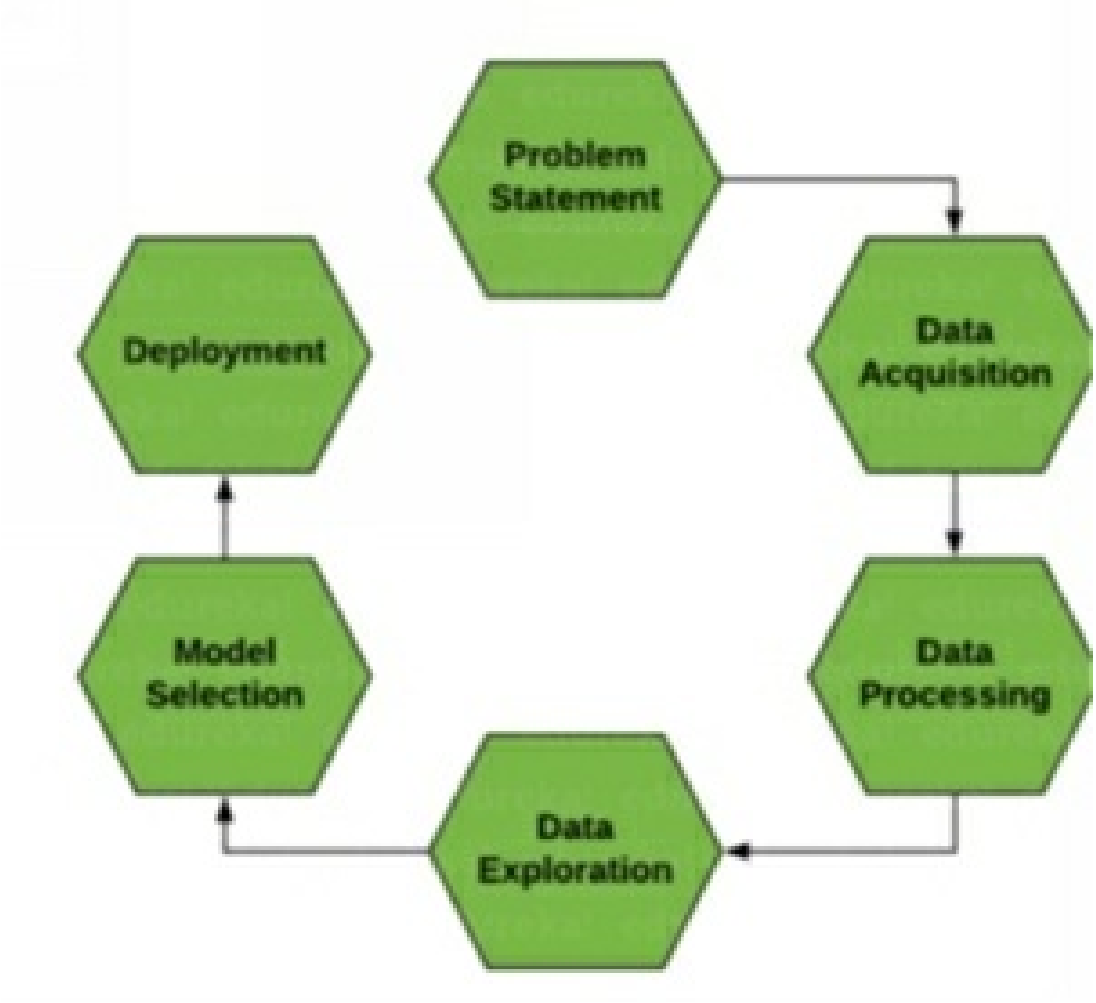
- **Linear or non-direct:** This element characterizes how the information objects are coordinated in a successive arrangement, similar to a rundown or an unordered grouping, similar to a table.
- **Homogeneous or non-homogeneous:** This capacity characterizes how all information objects in an assortment are similar or of various types.

- **Static or dynamic:** This procedure decides to show to gather the information structures. Static information structures at aggregation time have fixed sizes, structures, and objections in the memory. Dynamic information types have measurements, instruments, and objections of memory that may recoil or extend contingent upon the application.

Information Structure Types

Sorts of the information structure are dictated by what kinds of tasks will be required or what sorts of calculations will be executed. This incorporates:

- **Clusters:** An exhibit stores a rundown of memory things in nearby areas. Segments of a similar classification are found together since every component's position can be handily determined or gotten to. Exhibits can be fixed in size or adaptable long.
- **Stacks:** A stack holds a bunch of articles in straight requests added to activities. This request might be past due in first out (LIFO) or first-out (FIFO).
- **Lines:** A-line stores a stack-like determination of components; nonetheless, the movement's succession must be first in the first out. Connected records: In an explicit request, a connected rundown stores a choice of things. In a connected rundown, each unit or hub incorporates an information thing just as a source of perspective or connection to the following component in the rundown.
- **Trees:** A tree stocks a theoretical, progressive assortment of things. Every hub is associated with different hubs and can have a few sub-values, otherwise called a kid.
- **Charts:** A diagram stores a non-straight plan gathering of things. Charts comprise of a restricted arrangement of hubs, additionally called vertices, and lines associating them, otherwise called edges. They are valuable for portraying measures, all things considered, for example, organized PCs.
- **Attempts:** Atria or inquiry tree is regularly an information structure that stores strings as information records, which can be masterminded in a visual chart.
- **Hash tables:** A hash table or hash graph is contained in a social rundown that marks the keys to factors. A hash table uses a hashing calculation to change a file into various holders containing the ideal information. These information frameworks are called complex since they can contain huge amounts of related information. Instances of the base or major information structures are number, glide, Boolean, and character.



Usage of Information Structures

Information structures are commonly used to join the information types in actual structures. This can be deciphered into a wide scope of utilization, including a parallel tree indicating an information base table. Information structures are utilized in the programming dialects to coordinate code and data in computerized stockpiling. Python information bases and word references, or JavaScript clusters and items, are famous coding frameworks used to accumulate and investigate information. Likewise, information structures are an indispensable piece of a clear programming plan. Databases Data frameworks' noteworthiness is important to deal with immense volumes of information, for example, successfully, information put away in libraries, or ordering administrations.

The executives' precise information design requires memory portion identifier, information interconnections, and information measures, all of which uphold the information structures. Furthermore, it is critical to utilize information structures and choose the right information structure for every task.

Picking an inadmissible information structure could prompt moderate running occasions or messy code. Any contemplations that should be seen while picking an information framework incorporate what kind of data should be handled, where new information will be put, how the information will be coordinated, and how much space will be dispensed.

How Critical Is the Use of Python for Data Science?

- **Efficient and easy to utilize:** Python is viewed as an instrument for apprentices, and any understudy or analyst with just essential arrangement could begin chipping away at it. Time and cash spent troubleshooting codes and imperatives on various tasks the executives are additionally limited. The ideal opportunity for code execution is less contrasted with other programming dialects, such as C, Java, and C #, making designers and programming engineers invest unmistakably more energy taking a shot at their calculations.

- **Library Choice:** Python offers a massive library and AI and computerized reasoning information base.

- **Scalability:** It gives adaptability in tackling issues that can't be comprehended with other scripts. Numerous organizations use it to build up a wide range of quick methods and frameworks.

- **Visual Statistics and Graphics:** Python gives various perception instruments.

Python Data Science Uses

- **First stage:** First of all, we need to learn and comprehend what structure information takes. If we see information to be an enormous Excel sheet with sections and crows lakhs, at that point, maybe you should realize some solution for that. You need to accumulate data into each line just as a section by executing a few activities and looking for a particular kind of information. Finishing this sort of computational errand can burn-through a great deal of time and difficult work. Consequently, you can utilize Python's libraries, such as Pandas and NumPy, to rapidly finish the undertakings by

utilizing similar calculations.

- **The second stage:** The next obstacle is to get the information required. Since information isn't promptly available to us, we need to dump information from the organization varying. Here the Python Scrap and splendid Soup libraries can empower us to recover information from the web.
- **The third stage:** We should get the reenactment or visual introduction of the information at this progression. Driving points of view get troublesome when you have an excessive number of figures on the board. The right method to do is speak to the information in chart structure, diagrams, and different designs. The Python Seaborn and Matplotlib libraries are utilized to execute this activity.
- **Fourth stage:** The following stage is AI, which is greatly muddled registering. It incorporates numerical instruments, for example, the likelihood, analytics, and framework tasks of sections and lines over one hundred thousand. With Python's AI library Scikit-Learn, the entirety of this will turn out to be straightforward and powerful.

Chapter 11. Data Science and the Cloud

Data science is a mixture of many concepts. It is important to have some programming skills to become a data scientist. Even though you might not know all the programming concepts related to infrastructure, basic computer science concepts are a must. You must install the two most common and most used programming languages, i.e., R and Python, on your computer. With the ever-expanding advanced analytics, Data Science continues to spread its wings in different directions. This requires collaborative solutions like predictive analysis and recommendation systems. Collaboration solutions include research and notebook tools integrated with code source control. Data science is also related to the cloud. The information is also stored in the cloud. So, this lesson will enlighten you with some facts about the "data in the Cloud." So let us understand what cloud means and how the data is stored and how it works.

The Cloud

The cloud can be described as a global server network, each having different unique functions. Understanding networks is required to study the cloud. Networks can be simple or complex clusters of information or data.

Network

As specified earlier, networks can have a simple or small group of connected computers or large groups of computers. The largest network can be the Internet. The small groups can be home local networks like Wi-Fi and Local Area Network limited to certain computers or locality. There are shared networks such as media, web pages, app servers, data storage, printers, and scanners. Networks have nodes, where a computer is referred to as a node. The communication between these computers is established by using protocols. Protocols are the intermediary rules set for a computer. Protocols

like HTTP, TCP, and IP are used on a large scale. All the information is stored on the computer, but it becomes difficult to search for information on the computer every time. Such information is usually stored in a data Centre. Data Centre is designed so that it is equipped with support security and protection for the data. Since the cost of computers and storage has decreased substantially, multiple organizations opt to use multiple computers that work together that one wants to scale. This differs from other scaling solutions like buying other computing devices. The intent behind this is to keep the work going continuously. Even if a computer fails, the other will continue the operation. There is a need to scale some cloud applications, as well. Having a broad look at some computing applications like YouTube, Netflix, and Facebook requires some scaling. We rarely experience such applications failing, as they have set up their systems on the cloud. There is a network cluster in the cloud, where many computers are connected to the same networks and accomplish similar tasks. You can call it a single source of information or a single computer that manages everything to improve performance, scalability, and availability.

Data Science in the Cloud

The whole process of Data Science occurs in the local machine, i.e., a computer or laptop provided to the data scientist. The computer or laptop has inbuilt programming languages and a few more prerequisites installed. This can include common programming languages and some algorithms. The data scientist later has to install relevant software and development packages as per his/her project. Development packages can be installed using managers such as Anaconda or similar managers. You can opt for installing them manually too. Once you install and enter into the development environment, your first step, i.e., the workflow, starts where your companion is only data. It is not mandatory to carry out the task related to Data Science or Big data on different development machines. Check out the reasons behind this:

1. The processing time required to carry out tasks in the development environment fails due to processing power failure.
2. Check the presence of large data sets that cannot be contained in the development environment’s system memory.
3. Deliverables must be arrayed into a production environment and incorporated as a

large application component.

4. It is advised to use a machine that is fast and powerful.

Data scientists explore many options when they face such issues; they use on-premise machines or virtual machines that run on the cloud. Using virtual machines and auto-scaling clusters has various benefits, such as they can span up and discard it anytime in case it is required. Virtual machines are customized in a way that will fulfill one's computing power and storage needs. Deploying the information in a production environment to push it in a large data pipeline may have certain challenges. These challenges are to be understood and analyzed by the data scientist. This can be understood by having a gist of software architectures and quality attributes.

Software Architecture and Quality Attributes

Software Architects develop a cloud-based software system. Such systems may be a product or service that depends on the computing system. If you are building software, the main task includes selecting the right programming language that is to be programmed. The purpose of the system can be questioned; hence, it needs to be considered. Developing and working with software architecture must be done by a highly skilled person. Most organizations have started implementing effective and reliable cloud environments using cloud computing. These cloud environments are deployed over to various servers, storage, and networking resources. This is used in abundance due to its lower cost and high ROI.

The main benefit to data scientists or their teams is using the big space in the cloud to explore more data and create important use cases. You can release a feature and have it tested the next second and check whether it adds value or it is not useful to carry forward. All this immediate action is possible due to cloud computing.

Sharing Big Data In The Cloud

The role of Big Data is also vital while dealing with the cloud as it makes it easier to track and analyze insights. Once this is established, big data creates great value for users.

The traditional way was to process wired data. It became difficult for the team to share their information with this technique. The usual problems included transferring large amounts of data and collaboration of the same. This is where cloud computing started sowing its seed in the competitive world. All these problems were eliminated due to cloud computing, and gradually, teams were able to work together from different locations and overseas. Therefore, cloud computing is very vital in both Data Science as well as Big data. Most organizations make use of the cloud. To illustrate, a few companies that use the cloud are Swiggy, Uber, Airbnb, etc. They use cloud computing for sharing information and data.

Cloud And Big Data Governance

Working with the cloud is a great experience as it reduces resource cost, time, and manual efforts. But the question arises that how organizations deal with security, compliance, governance? Regulation of the same is a challenge for most companies. Big data problems are not limited, but working with the cloud also has its issues related to privacy and security. Hence, it is required to develop a strong governance policy in your cloud solutions. To ensure that your cloud solutions are reliable, robust, and governable, you must keep it an open architecture.

Need For Data Cloud Tools To Deliver High Value Of Data

The demand for data scientists in this era is increasing rapidly. They are responsible for helping big and small organizations develop useful information from the provided data or data set. Large organizations carry massive data that need to analyze continuously. As per recent reports, almost 80% of the organizations' unstructured data are in the form of social media, emails, i.e., Outlook, Gmail, etc., videos, images, etc. With the rapid growth of cloud computing, data scientists deal with various new workloads from IoT, AI, Blockchain, Analytics, etc. Pipeline. Working with all these new workloads requires a stable, efficient, and centralized platform across all teams. With all this, there is a need to manage and record new data and legacy documents.

Once a data scientist is given a task, and he/she has the dataset to work on, he/she must possess the right skills to analyze the ever-increasing volumes through cloud technologies. They need to convert the data into useful insights that would be responsible for uplifting the business. The data scientist has to build an algorithm and code the program. They mostly utilize 80% of their time to gather information, create and modify data, clean if required and organize data. Rest 20% is utilized for analyzing the data with effective programming. This calls for the requirement to have specific cloud tools to help the data scientist reduce their time searching for appropriate information. Organizations should make available new cloud services and cloud tools to their respective data scientists to organize massive data quickly. Therefore, cloud tools are very important for a data scientist to analyze large amounts of data in a shorter period. It will save the company time and help build strong and robust data models.

Conclusion

Almost everyone will agree with the statement that big data has arrived in a big way and has taken the business world by storm. But what is the future of data analysis, and will it grow? What are the technologies that will grow around it? What is the future of big data? Will it grow more? Or is the big data going to become a museum article soon? What is cognitive technology? What is the future of fast data? Let’s look at the answers to these questions. We’ll take a look at some predictions from the experts in data analysis and big data to get a clearer picture.

The data volume will keep on growing. There is practically no question in people's minds that we’ll keep developing a larger and larger quantity of data, especially after considering the number of internet-connected devices and handheld devices is going to grow exponentially. The ways we undertake data analysis will show marked improvement in the upcoming years. Although SQL will remain the standard tool, we’ll see other tools such as Spark emerging as a complementary method for the data analysis, and their number will keep on growing as per reports.

More and more tools will become available for data analysis, and some of them will not need the analyst. Microsoft and Salesforce have announced some combined features that will allow the non-coders to create apps to view the business data. The prescriptive analytics will get built into the business analytics software. IDC predicts that 50 percent of all software related to business analysis will become available with all the business intelligence it needs by the year 2020.

In addition to these features, real-time streaming insight into the big data will turn into a hallmark for the data winners moving forward. The users will be looking to use this data to make informed decisions within real-time by using Spark and Kafka programs. The top strategic trend that will emerge is machine learning. Machine learning will become a mandatory element for big data preparation and predictive analysis in businesses going forward.

You can expect big data to face huge challenges, especially in the privacy of

user details. The new private regulations enforced by the European Union intend to protect the personal information of the users. Various companies will have to address privacy controls and processes. It is predicted that most business ethics violations will be related to data in the upcoming years.

Soon you can pretty much expect all companies to have a chief data officer in place. Forrester says that this officer will rise in significance within a short period. Still, certain kinds of businesses and generation gaps might decrease their significance in the upcoming future. Autonomous agents will continue to play a significant role, and they will keep on being a huge trend, as per Gartner. These agents include autonomous vehicles, smart advisers, virtual personal assistants, and robots.

The staffing required for the data analysis will keep on expanding, and people from scientists to analysts to architects to the experts in data management will be needed. However, a crunch in big data talent availability might see the large companies develop new tactics. Some large institutes predict that various organizations will use internal training to get their issues resolved. A business model having big data in the form of service can be seen on the horizon.