**CHAPTER-1**

**INTRODUCTION**

Prediction is the forecasting about what will happen in the future, not always based on experience or knowledge but alarge amount of historical data and current data are gatheredand their behaviour is analysed. Predictive Analytic combine powerful, fully automated discovery and analysis technologies to enable decision maker to prepare for future by learning from past.

It helps banks to fetch the relevant data of customers, identify fraudulent activities, helps in application screening, capture relationships between predicted and explanatory variables from past happenings and uses it to predict future outcomes.

Retail banks have a tremendous opportunity to use insight extracted from data to refine their customer offering. These insights add a level of refinement and specifics at the customer, transaction and event levels and can help explain not only what happened, but why it happened and when, to which customer, and even potentially predict if it will occur again and under what circumstances?

A recent report by Ventana Research described price and revenue optimisation in banking as a natural fit for the application of big data analytics, sifting through large data sets to collect characteristics of consumer behaviour that enable banks to identify customer segments and quantify their price sensitivity.

The firm's senior vice president of research, Robert Kugel explains that big data analytics software can help users manage more granularity in the process of defining offers for customers (and the levels of discretion they allow to account managers and sales people to set prices) as well as the terms and conditions of the transaction.

Upon identifying characteristics that influence buyer's price sensitivity, banks can combine the most relevant factors to present a price that will enable them to optimise revenue and profits from those customers.

Enabling them to automate analytics and reporting as well as facilitating management of the related data, makes it easier for a bank to set prices in a way that best matches its strategic objectives, whether that is to be a market share leader in specific product categories or to maximise returns on certain assets.

Kugel observes that using price and revenue optimisation rather than simplistic risk-based pricing can provide a competitive advantage in achieving higher returns on assets and lower costs of capital. It also has the advantage of often being counterintuitive and therefore offering strategies unavailable to less well-informed competitors.

Predictive analytics allows banks to quantify the expected movements of balances in a way that eliminates some of the uncertainty. Being able to predict which customers are more price sensitive to acquisition and retention rates, enables the bank to plan for trade-offs and ultimately make better long term decisions.

References to the longer term also remind us that there are many potential applications of data analytics in retail banking that have yet to be developed. For example, analysis of customer spending behaviour patterns, exposures, and risks could eventually enable retail banks to offer services such as dynamic monthly credit card allowances.

The main expectations that banks have for data analytics include:

• Enabling them to tailor their offerings to the needs of individual customers

• Improving trading strategies

• Providing better insights into market dynamics and improve market research

• Improving their ability to react to internal and external issues

• Speeding up high quality decision-making processes

• Identifying possibilities for revenue enhancement and cost reduction

• Banks clearly recognise the need to get support in technology and data science to execute banking and customer strategies and many are partnering with us for this reason.

They are using such solutions for a variety of reasons: to predict future customer behaviour and thus tailor customer solutions; improve their bottom line; provide management insights for strategy solutions; provide answers to issues (including regulatory issues); make faster decisions; and enhance balance sheet and P&L management.

The value of this partnership is clear - we have helped these banking customers improve bottom line performance at the portfolio level by an average of 17%.

* 1. **Existing System:**

Banks are no longer places where people would go deposit or withdraw their money and that would be the end of the story. Banks irrespective of their size and span have diversified into many segments.

If one has to broadly classify the activities into categories, then they would mostly come under Retail and Corporate or Commercial Banking.



One of the ways that corporate banks can ensure profit is by sanctioning loans and other products of credit, which become a key factor in corporate banking. Managing the cash flow or the working capital of the corporate, with Treasury and currency conversion requirements is another area of function.

Corporate banks also structure and customise loans managing the expansion needs of businesses such as offering equipment loans for the manufacturing clientele. Trade finance, bill collection, credit letters and payroll are common tasks handled by this sector for the companies they are representing. Some banks might also offer commercial real estate analysis, portfolio evaluation, debt and equity structuring as well. Through the commercial banking intervention, banks might also offer asset management to their clients.

* 1. **Proposed System**:

Retail banking is generally the face of the bank situated in forms of various branches located across cities, who deal with daily retail customers, also mostly known as personal banking. These banks or branches purely deal with retail customer’s, their volumes are large in terms of clientele. You will not find many banks that purely focus on retail banking. It is interesting to note that retail banking represents the footing through which funds are collected for most banks.



There is a large number, a variety of products and services offered by most retail banks. Fundamental services include a quintessential checking and savings accounts, a line of credit like home mortgages, vehicle loans, personal loans at high-interest rates. Credit cards, remittance services are a few profitable offering that retail banking presents to its clients.

The financial strength of an individual is the deciding factor in the services that will be offered to the client. If the client is of high potential then the client might be offered a private banker, or a relationship manager who will manage their banking requirement, however, if the client is of modest means would need to perform the tasks themselves either through a teller or through the technology integrated mediums of transactions offered by most retail banks these days, like ATM, Online transactions, internet banking etc., to name a few.cz

* 1. **Software Requirements:**

**Python** is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990.

**Python** is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* **Python is Interpreted** − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive** − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* **Python is Object-Oriented** − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* **Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

## Characteristics of Python:

* It supports functional and structured programming methods as well as OOP.
* It can be used as a scripting language or can be compiled to byte-code for building large applications.
* It provides very high-level dynamic data types and supports dynamic type checking.
* It supports automatic garbage collection.
* It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

**Applications of Python:**

Python is one of the most widely used language over the web.

* **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
* **Easy-to-read** − Python code is more clearly defined and visible to the eyes.
* **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
* **A broad standard library** − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
* **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
* **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* **Extendable** − You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
* **Databases** − Python provides interfaces to all major commercial databases.
* **GUI Programming** − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
* **Scalable** − Python provides a better structure and support for large programs than shell scripting.

Jupyteris a comprehensive software suite for interactive computing, that includes various packages such as Jupyter Notebook, QtConsole, nbviewer, JupyterLab.

As a server-client application, the Jupyter Notebook App allows you to edit and run your notebooks via a web browser. The application can be executed on a PC without Internet access, or it can be installed on a remote server, where you can access it through the Internet.

Its two main components are the kernels and a dashboard.

A kernel is a program that runs and introspects the user’s code. The Jupyter Notebook App has a kernel for Python code, but there are also kernels available for other programming languages.

The dashboard of the application not only shows you the notebook documents that you have made and can reopen but can also be used to manage the kernels: you can which ones are running and shut them down if necessary.

Watson Studio provides you with the environment and tools to solve your business problems by collaboratively working with data. You can choose the tools you need to analyze and visualize data, to cleanse and shape data, to ingest streaming data, or to create and train machine learning models.

Node-RED is a browser-based flow editor that lets users wire together hardware devices, APIs, and online services in new and interesting ways.

Node-RED's nodes are like npm packages, and you can get them [the same way](http://flows.nodered.org/). And because Node-RED has a built-in text editor, you can make applications as complex as you like by adding JavaScript functions.

Because Node-RED is based on Node.js and takes advantage of the event-driven, non-blocking model, it can be run on low-cost hardware like the Raspberry Pi or in the cloud.

* 1. **Hardware Requirements:**

Laptop

Processor And Memory Features

|  |  |
| --- | --- |
| Processor Brand | * Intel |
| Processor Name | * Core i5 |
| Processor Generation | * 8th Gen |
| SSD | * Yes |
| SSD Capacity | * 256 GB |
| RAM | * 8 GB |
| RAM Type | * DDR4 |
| Processor Variant | * 8250U |
| Clock Speed | * 1.6 GHz with Turbo Boost Upto 3.4 GHz |
| Expandable Memory | * Upto 16 GB |
| RAM Frequency | * 2400 MHz |
| Cache | * 6 MB |
| Graphic Processor | * Intel Integrated UHD620 |

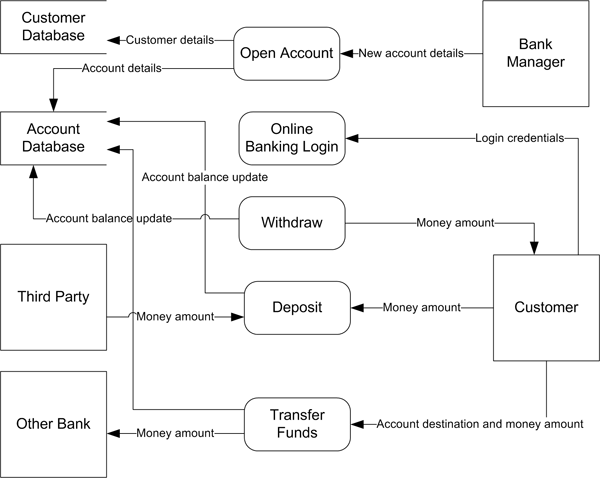
Operating System

|  |  |
| --- | --- |
| OS Architecture | * 64 bit |
| Operating System | * Windows 10 Home |

**CHAPTER-2**

**DESIGN OF THE PROJECT**

**FLOW CHART OF RETAIL BANKING:**

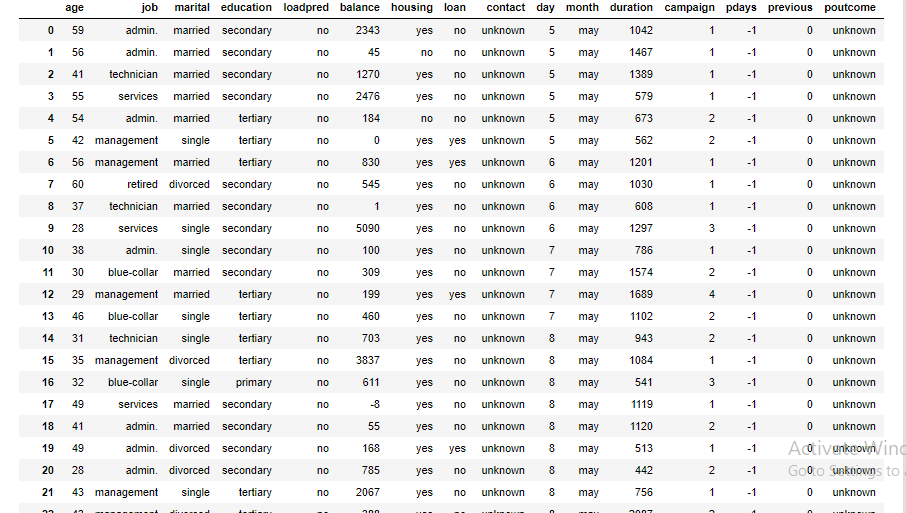


**Data Collection:**

In addition, to structured data, a large quantity of unstructured data is originating from the variety of relevant sources. The various data sources are Email, Call-Center,Social-Media, Websites, Customer feedbacks, agents and so on. This structured and unstructured data together combine into the large volume of business data that becomes useful in important decision making activity. Often, there is an immediate need to analyze data at hand, too discover pattern,reveals threats, monitor critical systems and make decisionsabout the direction the organization should take. Analysis of these data helps the organization in determining next path of action for the betterment of their organization as well as also in determining consumer profit.

The attributes we used in this predictive Analytics for Retail Banking are:

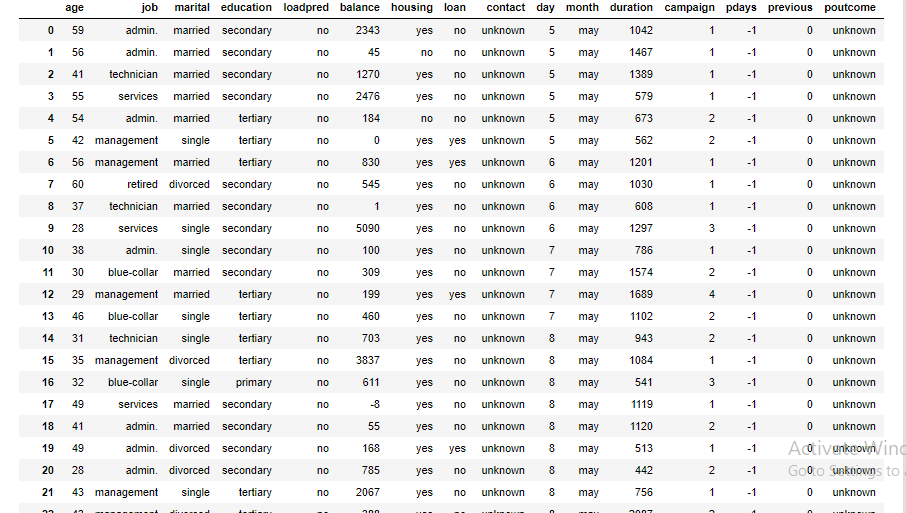
* Age
* Job
* Marital
* Education
* Loan Pred
* Balance
* Housing
* Loan
* Contact
* Month
* Duration
* Campaign
* Pdays
* Previous
* Poutcome
* Deposit

****

**CHAPTER-3**

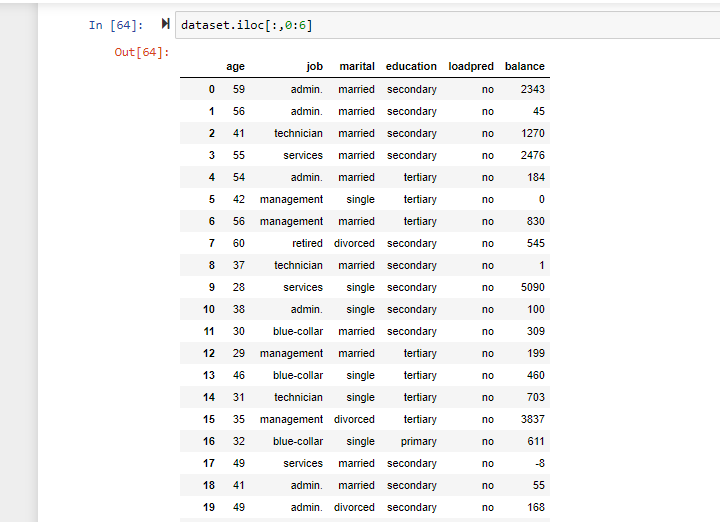
**IMPLEMENTATIONS**

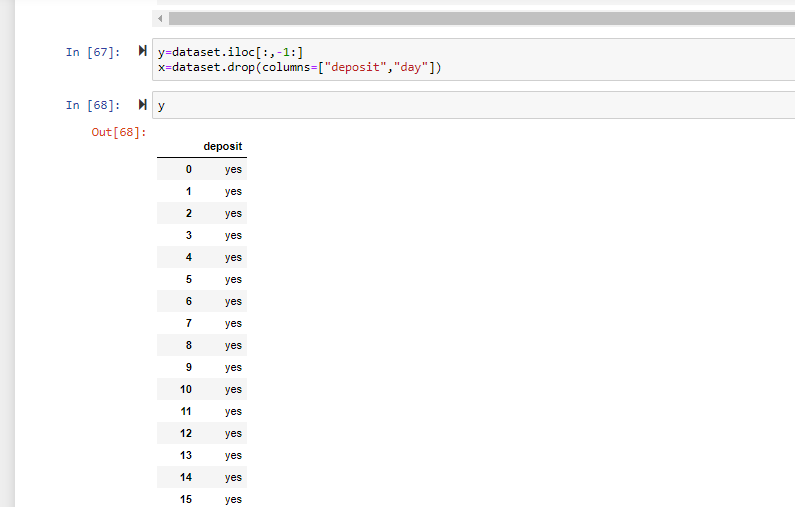
**Step1:** import libraries and dataset.Import the important libraries and the dataset we are using to perform Random Forest.

****

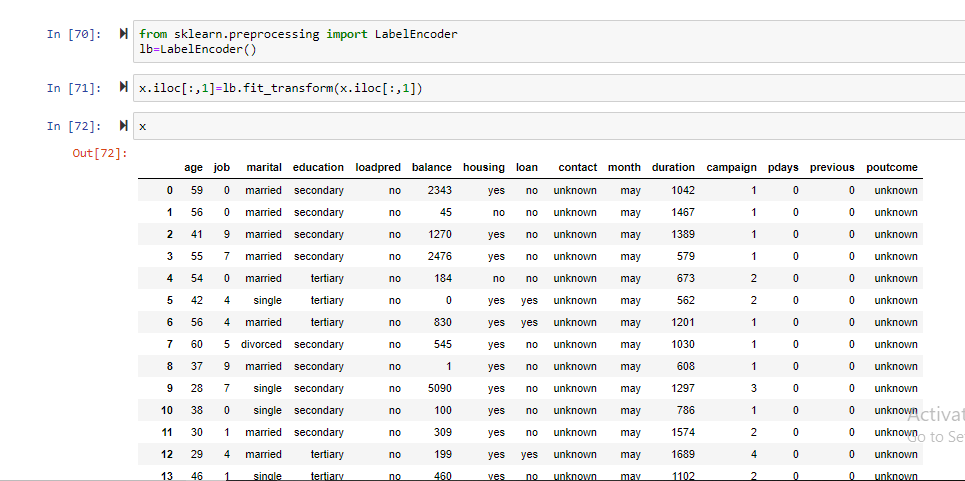
**Step2:** check for datatype and null values.

**Step 3:** Dividing the dataset into 2 componentsDivide dataset into two components that is X acnd y. X will contain the Column between 0 and 6. y will contain the 7column.(i.e (:,-1:)).

****

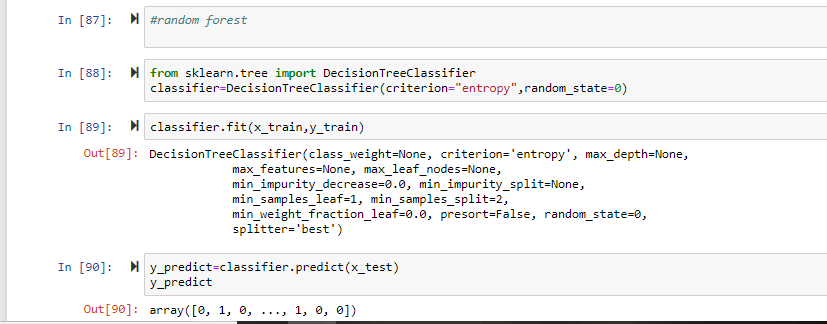
****

**Step 4:**Now to convert the following in to numerical values,we use Label Encoder .

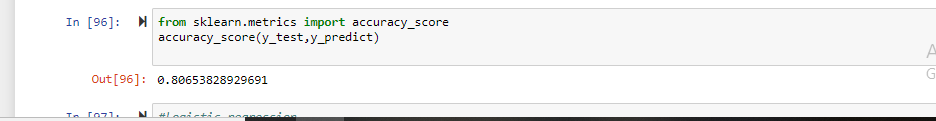
****

**Step 5:**Fitting Random Forest to the dataset

Fitting the Random Forest model On two components.

****

**Step 6:** calculate the accuracy\_score

****

**CHAPTER-4**

**TESTING**

As we have tried regression and classification algorithms ,we got highest accuracy score for random forest algorithm i.e related to classification.

**What is Random forest algorithm?**

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

If you know the decision tree algorithm. You might be thinking are we creating more number of decision trees and how can we create more number of decision trees. As all the calculation of nodes selection will be same for the same dataset.

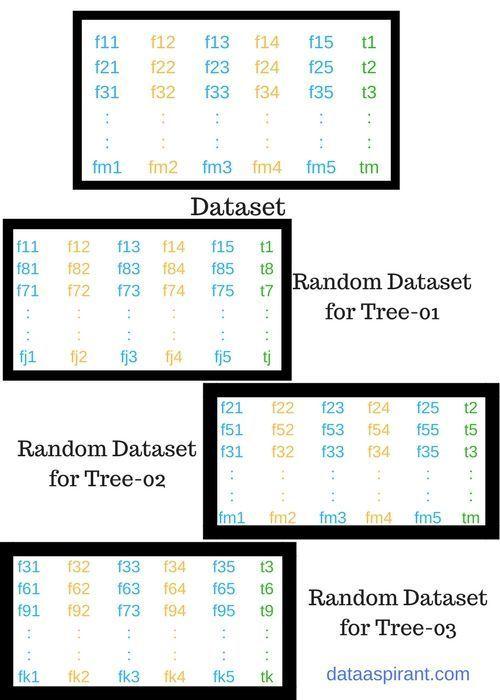
## Why Random forest algorithm?

To address why random forest algorithm. I am giving you the below advantages.

* The same random forest algorithm or the random forest classifier can use for both classification and the regression task.
* Random forest classifier will handle the missing values.
* When we have more trees in the forest, random forest classifier won’t overfit the model.
* Can model the random forest classifier for categorical values also.

**Random Forest Example**

## How Random forest algorithm works

****

The pseudocode for random forest algorithm and later we can walk through each step in the random forest algorithm.

The pseudocode for random forest algorithm can split into two stages.

* Random forest creation pseudocode.
* Pseudocode to perform prediction from the created random forest classifier.

First, let’s begin with random forest creation pseudocode

#### Random Forest pseudocode:

1. Randomly select “k” features from total “m” features.
   1. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

In the next stage, we are using the randomly selected “k” features to find the root node by using the [best split](https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/) approach.

The next stage, We will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node.

Finally, we repeat 1 to 4 stages to create “n” randomly created trees. This randomly created trees forms the random forest.

#### Random forest prediction pseudocode:

To perform prediction using the trained random forest algorithm uses the below pseudocode.

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm we need to pass the test features through the rules of each randomly created trees. Suppose let’s say we formed 100 random decision trees from the random forest.

Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Suppose the 100 random decision trees are prediction some 3 unique targets x, y, z then the votes of x is nothing but out of 100 random decision tree how many trees prediction is x.

Likewise for other 2 targets (y, z). If x is getting high votes. Let’s say out of 100 random decision tree 60 trees are predicting the target will be x. Then the final random forest returns the x as the predicted target.

This concept of voting is known as majority voting.

Now let’s look into few applications of random forest algorithm.

## Random forest algorithm applications

****

The random algorithm used in wide varieties applications. In this article, we are going address few of them.

Below are some the application where random forest algorithm is widely used.

1. Banking
2. Medicine
3. Stock Market
4. E-commerce

Let’s begin with the banking sector.

#### 1.Banking:

In the banking sector, random forest algorithm widely used in two main application. These are for finding the loyal customer and finding the fraud customers.

The loyal customer means not the customer who pays well, but also the customer whom can take the huge amount as loan and pays the loan interest properly to the bank. As the growth of the bank purely depends on the loyal customers. The bank customers data highly analyzed to find the pattern for the loyal customer based the customer details.

In the same way, there is need to identify the customer who are not profitable for the bank, like taking the loan and paying the loan interest properly or find the outlier customers. If

the bank can identify theses kind of customer before giving the loan the customer.  Bank will get a chance to not approve the loan to these kinds of customers. In this case, also random forest algorithm is used to identify the customers who are not profitable for the bank.

#### 2.Medicine

In medicine field, random forest algorithm is used identify the correct combination of the components to validate the medicine. Random forest algorithm also helpful for identifying the disease by analyzing the patient’s medical records.

#### 3.Stock Market

In the stock market, random forest algorithm used to identify the stock behavior as well as the expected loss or profit by purchasing the particular stock.

**DATA MODELING USING SUPERVISED MACHINE LEARNING TECHNIQUES:**

Python is a general-purpose high level programming language that is widely used in data science and for producing deep learning algorithms.

Deep structured learning or hierarchical learning or deep learning in short is part of the family of machine learning methods which are themselves a subset of the broader field of Artificial Intelligence.

Deep learning is a class of machine learning algorithms that use several layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.

Deep neural networks, deep belief networks and recurrent neural networks have been applied to fields such as computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, and bioinformatics where they produced results comparable to and in some cases better than human experts have.

Artificial Intelligence (AI) is any code, algorithm or technique that enables a computer to mimic human cognitive behaviour or intelligence.

Machine Learning (ML) is a subset of AI that uses statistical methods to enable machines to learn and improve with experience.

Deep Learning is a subset of Machine Learning, which makes the computation of multi-layer neural networks feasible. Machine Learning is seen as shallow learning while Deep Learning is seen as hierarchical learning with abstraction.

In supervised learning, we learn to predict values from labelled data. One ML technique that helps here is classification, where target values are discrete values; for example,cats and dogs.

In unsupervised learning, we make inferences from the input data that is not labelled or structured. If we have a million medical records and we have to make sense of it, find the underlying structure, outliers or detect anomalies, we use clustering technique to divide data into broad clusters.

Data sets are divided into training sets, testing sets, validation sets and so on.

A deep neural network (DNN) is an ANN with multiple hidden layers between the input and output layers. Similar to shallow ANNs, DNNs can model complex non-linear relationships.

### a. Recurrent Neuron

It’s one of the best from the [Deep Learning](https://data-flair.training/blogs/deep-learning/) Terminologies. Basically, in this output is sent back to the neuron for t timestamps. After looking at the diagram, we can say output is back as input t times. Also, we have to connect different together that will look like an unrolled neuron. Although, an important thing is that it provides us a more generalized output.

### b. RNN (Recurrent Neural Network)

We use a recurrent [neural network](https://data-flair.training/blogs/neural-network/), especially for sequential data. As in this, we use the previous output to predict the next one. Also, in this case, loops have a network within them. In a hidden neuron, loops have the capability to store information. As it stores previous words to predict the output.

### c. Vanishing Gradient Problem

It’s one of the best from the Deep Learning Terminologies. Where the activation function is very small, this problem arises. At the time of backpropagation, we have to multiply weights with low gradients. Although, they are small and vanish if they go further deep in the network. As for this reason, the neural network forgets the long-range dependence. Also, it becomes a problem of neural networks. As a result, dependence is very important for the network to remember.

We use activation function to solve problems like ReLu which do not have small gradients.

### d. Exploding Gradient Problem

We can say this is the opposite of the vanishing gradient problem. It is different as the activation function is too large. Also, it makes the weight of particular node very high. Although, we can solve it by clipping the gradient. So that it doesn’t exceed a certain value.

### e. Pooling

It’s one of the best from the Deep Learning Terminologies. We can introduce pooling layers in between the convolution layers. Basically, use this to reduce the number of parameters. Although, prevent over-fitting. Although, the size of the most common type of pooling layer of filter size(2,2) using the MAX operation. Further, we can say what it would do is, it would take the maximum of each 4\*4 matrix of the original image.

### f. Padding

In this process, we have to add an extra layer of zeros across the images. So, output image has the same size as the input. Hence, called as padding. If pixels of the image are actual or valid, we can say it’s a valid padding.

### g. Data Augmentation

It refers to the addition of new data that come from the given data, which might prove to be beneficial for prediction.

### h. Softmax

We use softmax activation function in the output layer for classification problems. It’s like sigmoid function. Also, the difference is that outputs are normalized, to sum up to 1.

It is like the sigmoid function, with the only difference being that the outputs are normalized, to sum up to 1. The sigmoid function would work in case we have a binary output. But we also have a multiclass classification problem. In this process softmax makes it easy to assign values to each class. Also, that can be interpreted as probabilities.

### i. Neural Network

[Neural Network](https://data-flair.training/blogs/learning-rules-in-neural-network/) form the backbone of deep learning. The goal of it is to find an approximation of an unknown function. It is a combination of interconnected neurons. These neurons have weights. Also, have a bias that needs to be updated during the network training depending upon the error. The activation function puts a nonlinear transformation to the linear combination. Thus, generates the output. The combinations of the activated neurons give the output.

### j. Input layer/ Output layer / Hidden layer

It’s one of the best from the Deep Learning Terminologies. The input layer is the one which receives the input. Also, it’s the first layer of the network. The output layer is the final layer of the network. These layers are the hidden layers of the network. We use these hidden layers to perform tasks on incoming data. Hence, pass generated output to the next layer. Although, both layers are visible but the intermediate layers are hidden.

### k. MLP (Multi-Layer perceptron)

We can not perform highly complex tasks by a single neuron. Therefore, we use stacks of neurons to generate the desired outputs. In the simplest network, we would have an input layer, a hidden layer, and an output layer. As in this, each layer has multiple neurons. Also, in each layer, all neurons are connected to all the neurons in the next layer. These networks are fully connected networks.

### l. Neuron

As we can say that we use neuron to form the basic elements of a brain. Also, helps to form the basic structure of a neural network. As we get new information. We start to generate an output.

### m. Weights

As soon as the input enters the neuron, we have to multiply it by a weight.

### n. Bias

We have to add another linear component to input in addition to weight, this is a bias. In input, we have to add weight multiplication. Basically, we have to add bias to change the range of the weight multiplied input. As soon as bias is added result will look like a\*W1+bias. Hence, it’s a linear component of the input transformation.

### o. Activation Function

As soon as we apply linear component to the input, a non-linear function is applied to it.  As this is done by applying the activation function to the linear combination. Hence, this translates the inputsignalstooutputsignals.The output after application of the activation function would look something like f(a\*W1+b) where f() is the activation function.

**p. Gradient Descent**

We use this as optimization algorithm for minimizing the cost.

Mathematically, to find the local minimum of a function one takes steps proportional to the negative of the gradient of the function.

### q. Learning Rate

We can say it is the amount of minimization in the cost function in each iteration. Also, one must be careful while choosing the learning rate. Since it should neither be very large that the optimal solution is missed. Also, not should be very low that it takes forever for the network to converge.

### r. Backpropagation

In back-propagation, the movement of the network is backward, the error along with the gradient flows back from the out layer through the hidden layers and updating of weights is done.

### s. Batches

In case of training of neural network, we divide in input into several chunks of equal size random. Instead of sending the entire input in one go. Also, trained data batches make the model more generalized.

### t. Epochs

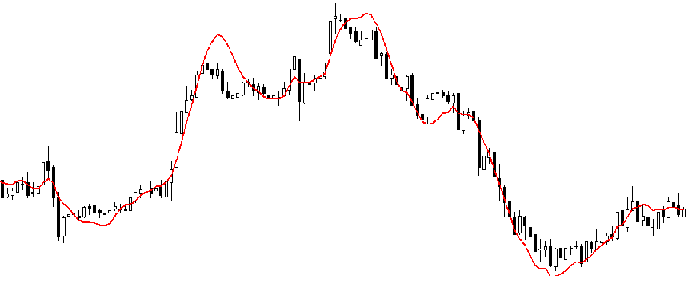
We can define it as a single training iteration. Then we define in term with batches in forwarding and backpropagation. This means 1 epoch is a single forward and backward pass of the entire input data.

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).

This technique is used for forecasting, time series modelling and finding the [causal effect relationship](https://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/) between the variables.

**REGRESSION:**

For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

****

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.  I’ll explain this in more details in coming sections.

**Why do we use Regression Analysis?**

As mentioned above, regression analysis estimates the relationship between two or more variables. Let’s understand this with an easy example:

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

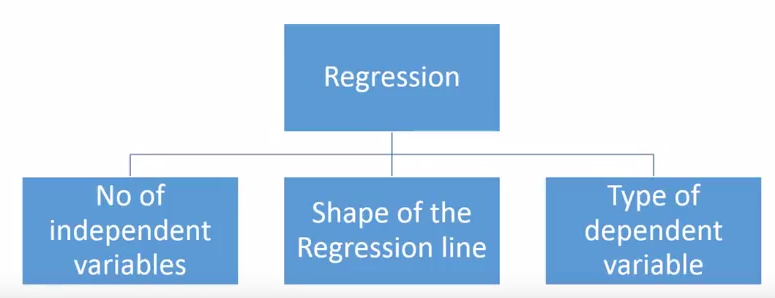
There are multiple benefits of using regression analysis. They are as follows:

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

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

**How many types of regression techniques do we have?**

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We’ll discuss them in detail in the following sections.

****

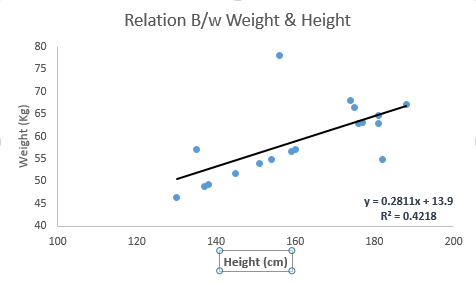
For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven’t used before. But before you start that, let us understand the most commonly used regressions:

**1. Linear Regression:**

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

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

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

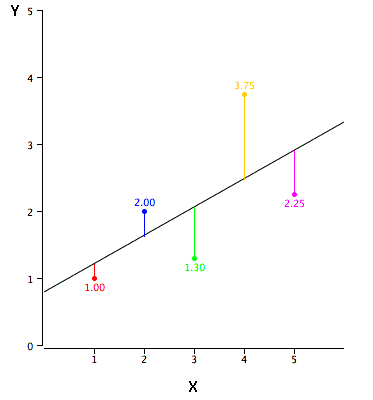
****

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.  Now, the question is “How do we obtain best fit line?”.

**How to obtain best fit line (Value of a and b)?**

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

**least square, regression line**

****

We can evaluate the model performance using the metric R-square.

**Important Points:**

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

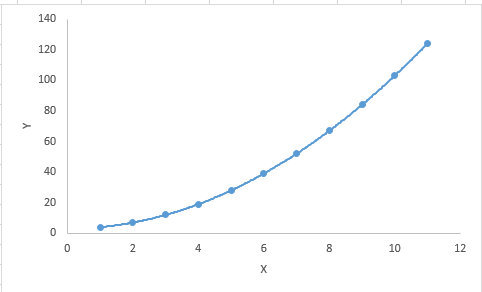
1. **Polynomial Regression:**

It is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x).

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

y=a+b\*x^2

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

****

1. **Logistic Regression:**

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

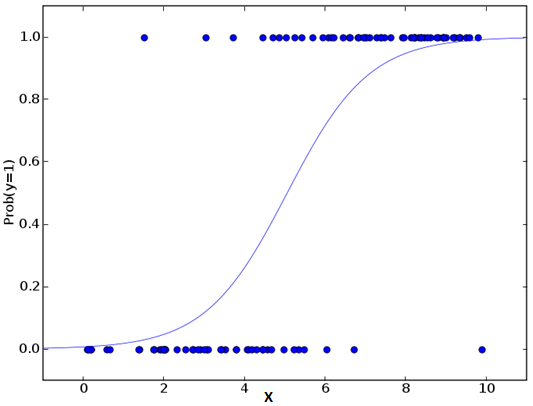
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is [logit](https://en.wikipedia.org/wiki/Logistic_function) function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

****

**Important Points:**

* It is widely used for classification problems
* Logistic regression doesn’t require linear relationship between dependent and independent variables.  It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
* To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
* It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
* The independent variables should not be correlated with each other i.e. no multi collinearity.  However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
* If the values of dependent variable is ordinal, then it is called as Ordinal logistic regression
* If dependent variable is multi class then it is known as Multinomial Logistic regression.

**3.MULTIPLE LINEAR REGRESSION :**

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable usually denoted by Y and a series of other changing variables known as independent variables. Regression model which contain more than two predictor variables are called Multiple Regression Model.

Multiple regression model is of the form: Y=b0+b1x1 +b2x2 +b3x3+ b4x4+...e

where, b0,b1,b2,b3,b4 are regression coefficient e is unexplained portion of dependent variable with zero mean and constant variance.

Multiple regression fits a model to predict a dependent (Y) variable from two or more independent (X) variables.

We can also use other algorithms like:

1.Decision tree

2.RandomForest

3.KNN(K- Nearest Neighbours).

**CLASSIFICATION:**

Classification predictive modeling problems are different from regression predictive modeling problems.

* Classification is the task of predicting a discrete class label.
* Regression is the task of predicting a continuous quantity.

There is some overlap between the algorithms for classification and regression; for example:

* A classification algorithm may predict a continuous value, but the continuous value is in the form of a probability for a class label.
* A regression algorithm may predict a discrete value, but the discrete value in the form of an integer quantity.

Some algorithms can be used for both classification and regression with small modifications, such as decision trees and artificial neural networks. Some algorithms cannot, or cannot easily be used for both problem types, such as linear regression for regression predictive modeling and logistic regression for classification predictive modeling.

Importantly, the way that we evaluate classification and regression predictions varies and does not overlap, for example:

* Classification predictions can be evaluated using accuracy, whereas regression predictions cannot.
* Regression predictions can be evaluated using root mean squared error, whereas classification predictions cannot.

Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation.

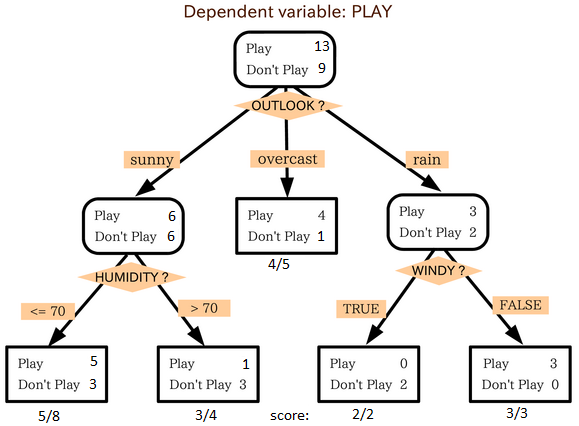
For example, an email of text can be classified as belonging to one of two classes: “spam“and “not spam“.

* A classification problem requires that examples be classified into one of two or more classes.
* A classification can have real-valued or discrete input variables.
* A problem with two classes is often called a two-class or binary classification problem.
* A problem with more than two classes is often called a multi-class classification problem.
* A problem where an example is assigned multiple classes is called a multi-label classification problem.

It is common for classification models to predict a continuous value as the probability of a given example belonging to each output class. The probabilities can be interpreted as the likelihood or confidence of a given example belonging to each class. A predicted probability can be converted into a class value by selecting the class label that has the highest probability.

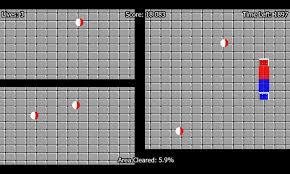
## 1. Decision Tree:

## It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

****

In the image above, you can see that population is classified into four different groups based on multiple attributes to identify ‘if they will play or not’. To split the population into different heterogeneous groups, it uses various techniques like Gini, Information Gain, Chi-square, entropy.

The best way to understand how decision tree works, is to play Jezzball – a classic game from Microsoft (image below). Essentially, you have a room with moving walls and you need to create walls such that maximum area gets cleared off without the balls.

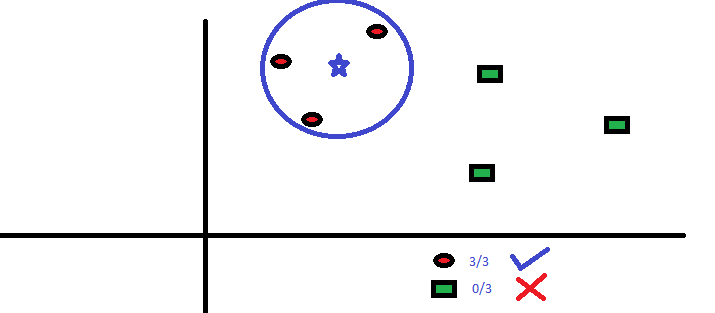


So, every time you split the room with a wall, you are trying to create 2 different populations with in the same room. Decision trees work in very similar fashion by dividing a population in as different groups as possible.

**2. kNN (k- Nearest Neighbors):**

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function.

These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modeling.

****

KNN can easily be mapped to our real lives. If you want to learn about a person, of whom you have no information, you might like to find out about his close friends and the circles he moves in and gain access to his/her information!

**Things to consider before selecting kNN:**

* KNN is computationally expensive
* Variables should be normalized else higher range variables can bias it
* Works on pre-processing stage more before going for kNN like outlier, noise removal.

**3. Random Forest:**

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we’ve collection of decision trees (so known as “Forest”). To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is planted & grown as follows:If the number of cases in the training set is N, then sample of N cases is taken at random but with replacement. This sample will be the training set for growing the tree.

* If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
* Each tree is grown to the largest extent possible. There is no pruning.

**4.POLYNOMIAL LINEAR REGRESSION**:

 polynomial regression is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) in which the relationship between the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) *x*  and the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) *y* is modelled as an *n*th degree [polynomial](https://en.wikipedia.org/wiki/Polynomial) in *x*.

Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of *y*, denoted E(*y* |*x*), and has been used to describe nonlinear phenomena such as the growth rate of tissues, the distribution of carbon isotopes in lake sediments, and the progression of disease epidemics.

 Although polynomial regression fits a nonlinear model to the data, as a [statistical estimation](https://en.wikipedia.org/wiki/Estimation_theory) problem it is linear, in the sense that the regression function E(*y* | *x*) is linear in the unknown [parameters](https://en.wikipedia.org/wiki/Parameter) that are estimated from the [data](https://en.wikipedia.org/wiki/Data).

For this reason, polynomial regression is considered to be a special case of [multiple linear regression](https://en.wikipedia.org/wiki/Multiple_linear_regression).

Polynomial regression models are usually fit using the method of [least squares](https://en.wikipedia.org/wiki/Least_squares).

The least-squares method minimizes the [variance](https://en.wikipedia.org/wiki/Variance) of the [unbiased](https://en.wikipedia.org/wiki/Bias_of_an_estimator) [estimators](https://en.wikipedia.org/wiki/Estimation_theory) of the coefficients, under the conditions of the [Gauss–Markov theorem](https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem).

The least-squares method was published in 1805 by [Legendre](https://en.wikipedia.org/wiki/Adrien-Marie_Legendre) and in 1809 by [Gauss](https://en.wikipedia.org/wiki/Gauss).

The first [design](https://en.wikipedia.org/wiki/Optimal_design) of an [experiment](https://en.wikipedia.org/wiki/Design_of_experiments) for polynomial regression appeared in an 1815 paper of [Gergonne](https://en.wikipedia.org/wiki/Joseph_Diaz_Gergonne" \o "Joseph Diaz Gergonne).

 In the twentieth century, polynomial regression played an important role in the development of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), with a greater emphasis on issues of [design](https://en.wikipedia.org/wiki/Design_of_experiments) and [inference](https://en.wikipedia.org/wiki/Statistical_inference).

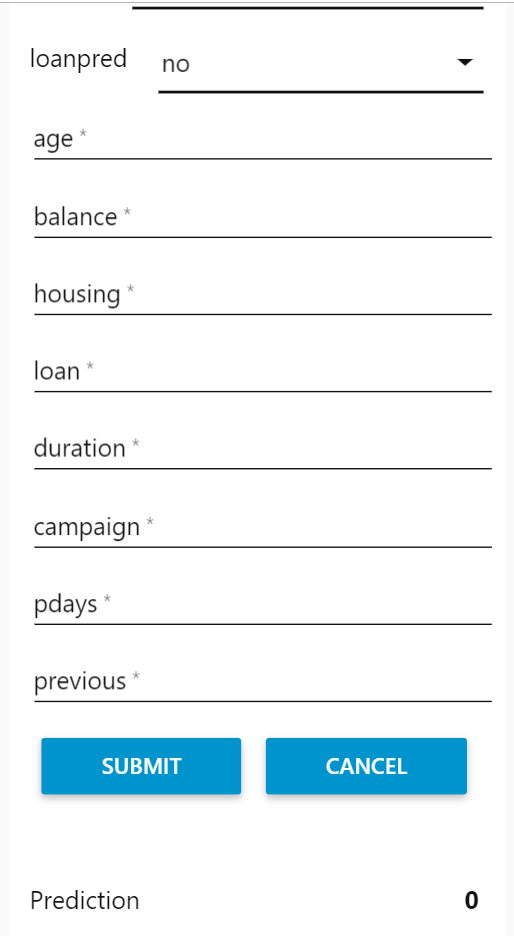
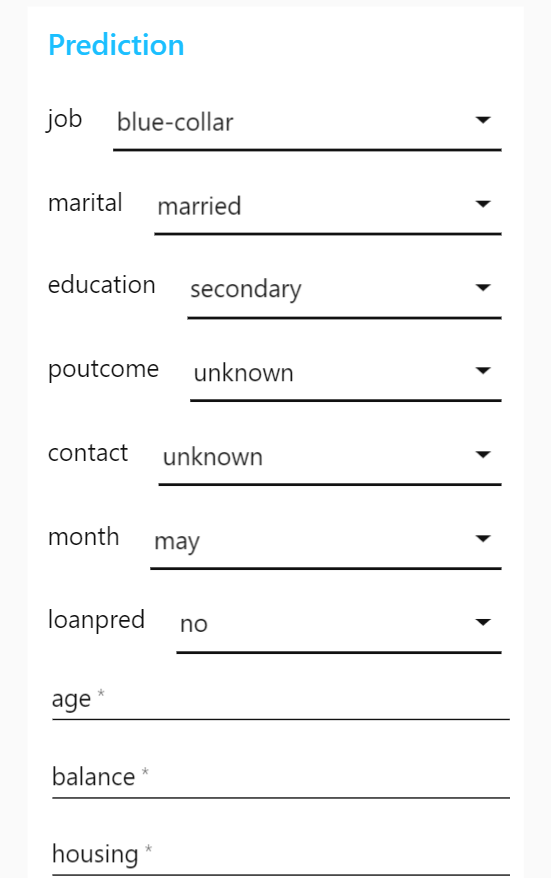
 More recently, the use of polynomial models has been complemented by other methods, with non-polynomial models having advantages for some classes of problems.

The explanatory (independent) variables resulting from the polynomial expansion of the "baseline" variables are known as higher-degree terms. Such variables are also used in [classification](https://en.wikipedia.org/wiki/Statistical_classification) settings.

* There are some relationships that a researcher will hypothesize is curvilinear. Clearly, such type of cases will include a polynomial term.
* Inspection of residuals. If we try to fit a linear model to curved data, a scatter plot of residuals (Y axis) on the predictor (X axis) will have patches of many positive residuals in the middle. Hence in such situation it is not appropriate.
* An assumption in usual multiple linear regression analysis is that all the independent variables are independent. In polynomial regression model, this assumption is not satisfied.

**CHAPTER-5**

**RESULT**



**CHAPTER-6**

**CONCLUSION**

Understanding the importance of analytics is very important. Banks should understand its potential, and how its power which is limitless can be utilized to the fullest. When using analytics, banks should bear in mind the needs of the customer, which is the base or the core of the analytics to be successful. There is a wide range of data generated everyday which goes untapped from various sources such as emails, social media, CRM and so on. With analytics, retail banks can use this data to drive key business decisions and deliver proactive services to the customers and also formulate business strategies. It has been observed that the use of analytics has been enhanced in the last decade; many companies are using analytics to satisfy their customer, retain them and also make decision about the right product to right customer at right time so as to increase the financial standings of the companies.

**FUTURE SCOPE:**

**PREDICTIVE ANALYSIS HELPS IN BANKING:**

**1. Fraud Detection**

Analytics can be used to recognize frauds that are not very obvious and then predictive analytics can be implemented on them to analyze them further. Data integration, utilizing unstructured data and machine learning techniques like supervised and unsupervised learning can help detect fraud cases.

**2. Application screening**

Predictive analysis in banking can help process huge volumes of applications,without excluding important variables, without delays or errors, without growing tired- all of it with regularity and steadiness. The results are very much accurate and authentic to be used. So ultimately there doesn't have to be a comparison between traditional or manual screening and analytics.

**3. Customer acquisition & retention**

Predictive analytics help in the process for optimized targeting, making it easier for banks to instantly identify the high-value customer segments most likely to respond. The customer base can further expand by acquiring the right type of customer.

predictive analytics helps banks and financial institutions retain their customers. It Identifies the customers most likely to defect before they end their relationship.Keeps the right customers longer.Predicts which actions will earn their loyalty.

**4. Knowing customer buying habits**

Targeting the right product and tracking customer usage is a challenge before banks which they independently and in conjunction with retailers are trying to curb.

With predictive analytics, banks can rapidly segregate various customer segments and replace it with highly relevant, individualized messages tailored to each customer’s profile, resulting in a higher response rate.

**5. Cross-selling**

Efficient cross-selling of products can happen by analyzing the existing customer behavior at places where multiple products are offered. Which specific products are to be sold to whom hence predicting the outcome is what successful cross-sellers do. And all of this results in more effective cross-selling thus increasing profitability and strengthening the customer relationship. Today, securing one profitable customer is a big task for banks, hence cross-selling another product to an existing customer helps a lot.Predictive analytics helps examine customers’ usage, spending, and other behavior and leads to effective cross-selling of the right product at the right time.

**6. Collections**

Banks have a mix of customers who always pay on time and those who lag.It is a tricky task to keep a track and maintain records of all individuals and differentiate who to focus more.

**7. Marketing optimization**

Predictive analytics help marketers to plan marketing campaigns and programs and monitor the results closely. By providing an insight into customer behavior and attitudes, and a complete, current view of your customers, analytics help your marketing team deliver the right message at the right time to the right customers.

**8. Customer Lifetime Value (LTV)**

Customer's lifetime value is how long the organizations are able to retain their customers. Know which customers should be the focus of new customer engagement efforts.Identify the previous factors that enhanced returns on customer engagements in the past.Use that knowledge to understand why customers responded to certain messages and promotions.

**9. Feedback management**

Feedback management is really important. Predictive analytics allows banks and financial firms to keep up their relationship with the customers by giving them the right services and products for their need and matching individual preferences in the most sorted way.

**FINDINGS AND SUGGESTIONS**

Economic prosperity has changed consumer demographics and technological factors have made consumers demand for better quality and efficient services. The service industry is becoming major contributor to the economy of many countries which were earlier dependent on the manufacturing sector. Service industry particularly the banking sector is not left behind in the competition. Banking industry has been highly commoditized. To be in business, every retail bank should have competitive differentiation that can be realized to a great extent through customer service excellence. The aim of retail banking industry is to satisfy customers and deepen their relationships. This can be achieved by taking the benefits of every cross-selling and up-selling opportunity. The availability of advanced technologies will help in boosting the cross-selling, increase customer retention and differentiate the brands in the retail banking sector.

The retail banking industry is facing stiff competition and the current scenario is that of the survival of the fittest. All the banks are trying to expand their customer base and are developing their own long-term strategies to stay in the market. To improve the customer services and relationship management many of the retail banks adapt information technology that has helped in integrating and consolidating banking operation. Competition in retail banking is increasing between the existing players, from international players and even from the new entrants. Retail banks should cope with the competition by providing excellent service through customer orientation for which can be achieved through specialization.

Retail banks in India are designing specialized are customized products keeping in mind their customer segment. To be able to up-sell and cross-sell their products, retail banks have to keep a good relationship with their client segment. The client segment is categorized according to the location of retail banks too like the urban and rural areas. Strategies are developed differently for different segments. Retail banks implement segment specific channel strategies to develop highperformance by migrating clients to cost-effective direct channel. Another strategy adopted by retail banks are the development of contact centre services and processes for high-end and low-end customers. Retail banks are increasing their cross-selling and up-scaling activities for private banking clients with the intention of increasing their customer base and improve their customer relation. Product innovation is another strategy applied by retail banks. The different client segments are offered other services like insurance or leasing services.

The latest strategy is in the use of debit/ATM cards in all processing platforms irrespective of the retail banks. In other words a single credit/debit/ATM card can be used in any of the ATM machine without any processing or transaction fee. Retail banks adopted the strategy of strengthening the online offerings by implementing new direct net features and also integrating the newer and the traditional sales channels for the clients who are tech-savvy and use the internet for their transactions and dealing in the banks. Retail banks adopt the strategy of increasing their product penetration to the existing clients in the traditional market while for the urban or metro markets, increasing the distribution and selling of specialized business products to commercial customers is focused more. Many banks focus on those markets that provide them with the best mix market growth and target clients concentration.

The broad objectives of the research were to find out the compatibility between objective and Service Quality of Retail Banking in India. The study also aimed at ascertaining the gap in perception of customers on the Service Quality Dimension between Public Sector Banks and Private Sector Banks. The hypotheses constructed have been tested by means of the survey conducted to elicit the perception of customers on the Service Quality dimension (Tangibles, Reliability, Responsiveness, Assurance and ATM Service Quality) between Public Sector Banks and Private Sector Banks. The analysis of the data collected through the survey has few positive results and confirmed the hypothesis of the study.

**BIBLIOGRAPHY**

Bibliography:<https://www.kaggle.com/sharanmk/bank-marketing-term-deposit>