[CHAPTER – 1]: INTRODUCTION TO COMPANY

ASG Technologies Pvt. Ltd.

We work with Companies in building Data Science based solutions for sales, Operations and Finance. Analytics and Data Science with Functional and Technology enabled solutions in Analytics space.
Our Industry driven Analytics and Data Science Internship program is crafted after intensive study of company needs.

We provide strategy consulting, managed services and support on Analytics, MDM, Data Governance, Big Data, Visualization and Cloud. Our Data Scientists have helped number of financial institutions, high-tech and Saas companies establish analytics practice to drive their growth via new products, services, partners and markets. In addition, we also specialize in Hadoop, SAP HANA/BI, MDM, Tableau and Alteryx implementations.

We also specialize in transforming credit unions and banks, take advantage of data and digital to better identify, target and engage with their customers. As the market faces pressure from FinTech innovations, digital disruption, constantly-evolving regulatory compliance and higher customer expectations, we help them to be more data driven and digitally ready for new opportunities.

ASGT LABS, a part of Decision Minds has several intellectual property and training programs to help organizations fast track analytics and cloud implementations.

Our experts across U.S. and ISO Certified (Information Security) low cost facilities, combine leadership in this space with newer delivery models (Agile/DevOps) to achieve higher speed, ownership, quality, and cost savings.

- ASG Technologies Pvt. Ltd.
- 501, C-5, BSI Business Park
- Sector-62, Noida - 201301

[CHAPTER – 2]: INTRODUCTION TO POJECT

2.1 Overview

Customer Churn Prediction

What is Churn?

Churn means (loss of customers to competition)

Accurately predicting customer churn using large scale time-series data is a common problem facing many business domains. The creation of model features across various time windows for training and testing can be particularly challenging due to temporal issues common to time-series data. In particular, we describe an effective method for handling temporally sensitive feature engineering.

Today numerous companies are prompt all over the world. Telecommunication market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers. Churn is the activity of the industry is the customers leaving the current company and moving to another telecom company. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up with the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning. In the telecommunication industry each company provides the customers with huge incentives to attract them to switch to their services, it is one of the reasons that customer churn is a big problem in the industry nowadays.

To prevent this, the company should know the reasons for which the customer decides to move on to another company. The information available can be analysed in different perspectives to provide various ways to the operators to predict and reduce churning. Only the relevant details are used in the analysis which contributes to the study from the information given. Data mining techniques are used for discovering the interesting patterns within data and it helps to learn to predict whether a customer will churn or not based on customer's data stored in the database.

2.2 Existing System

The majority of previous work in churn prediction has focused on predicting with a fixed set of categories. The developed models often rely on hard-coded concepts and sentence templates, which imposes limits on their variety. However, while closed vocabularies of prediction concepts constitute a convenient modeling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose. Moreover, the focus of these works has been on reducing complex categories into a

single feature, which can be considered to be an unnecessary restriction. Also, most previous attempts tried to handle the two problems discussed in section independently and stitch together the two parts to go from features to usable classifiers. Previous works tried to predict customer satisfaction within the model and get labels(columns) using them, and then try to make the best possible (prediction) out of the features obtained.

2.3 User Requirement Analysis

Requirements are broadly classified into two categories, that are Functional requirements and Non-Functional requirements. Functional requirements are the ones that are required for the functioning of the project and the Non-Functional requirements are the one's which are used for its training but are not playing any part further in working of the project.

Functional Requirement

It requires following tools for implementation:

Python (Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.)

Numpy(NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.)

Non-Functional Requirements

Performance (Performance of network depends on the strength of wifi connection, network used and also available machine for running model. It is expected to run a lot faster in same machine as that of two different machines exchanging messages.)

2.3.1 Data requirement

We are having 10 datasets in csv format named:

1. AccountMetrics.csv
2. AllCustomerChurnIndicator.csv
3. CommunityMetrics.csv
4. EngagementSurveyMetrics.csv

5. ITSAAvgPrice.csv
6. KnowledgeEventMetrics.csv
7. LicensedAdoptionMetrics.csv
8. NPSMetrics.csv
9. SupportSurveyMetrics.csv
10. UsageIndicator.csv

And finally we have done data mining and extraction the following datasets and derived a single dataset, with necessary value required our prediction, named as 'FinalDataset.csv'

2.3.2 Validation

The validation in the project basically refers to the testing of the project. The evaluations of Neural Style Transfer algorithm remain an open and important problem in this field. In general, there are two major types of evaluation methodologies that can be employed in this field, i.e., qualitative evaluation and quantitative evaluation. Qualitative evaluation relies on the aesthetic judgements of observers. The evaluation results are related to lots of factors (e.g., age and occupation of participants). While quantitative evaluation focuses on the precise evaluation metrics, which include time complexity, loss variation, etc.

2.3.3 Expected Hurdles

- Improper knowledge of the project.
- Improper knowledge of frameworks and languages.
- Absence of GPU.

2.3.4 SDLC model to be used

SDLC model used in this project is Waterfall Model. In the waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases. The waterfall Model illustrates the software development process in a linear sequential flow. This means that any phase in the development process begins only if the previous phase is complete. In this waterfall model, the phases do not overlap.

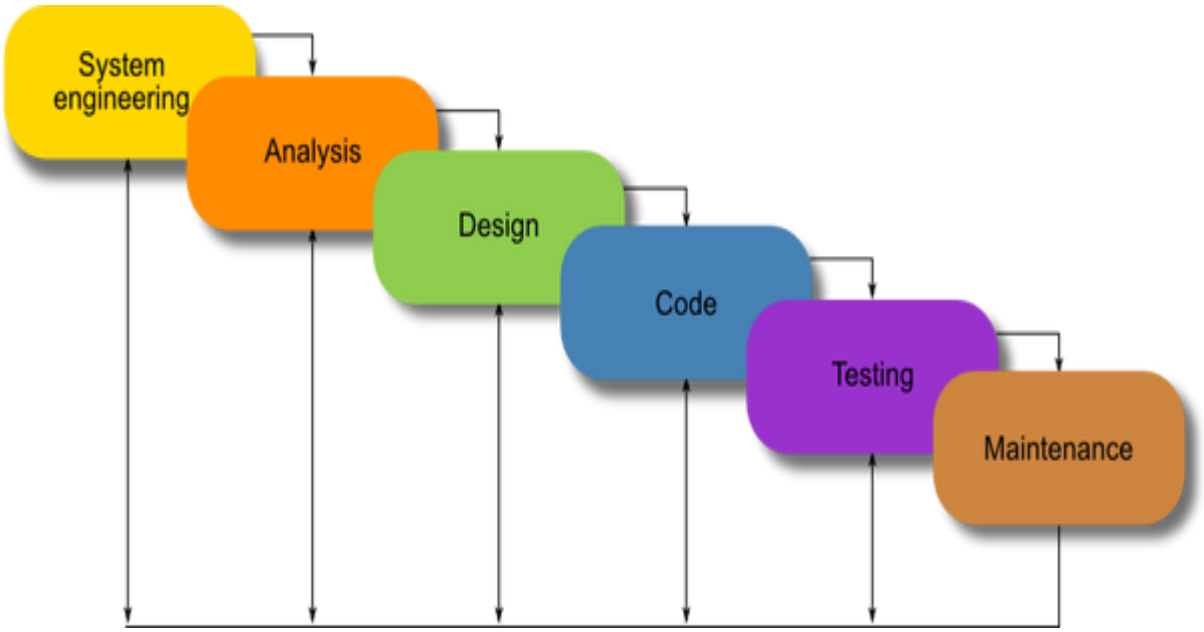


Fig.2.5.1 **Waterfall Model**

2.4 Feasibility study

From the inception of ideas for webapp system, until it is implemented and delivered to customer and even after that the system undergoes gradual developments and evaluations. The software is said to have life cycle composed of several phases. The webapp is said to have life cycle composed of several phases. At the feasibility stage , it is desirable that two or three different configuration will be pursed that satisfy the key technical requirement but which represent different level of ambition and cost. Feasibility is the determination of whether or not a project is worth doing. A feasibility study is carried out select a best system that make performance requirements. The data collected during primary investigation examines system feasibilities that is likelihood that the system will be beneficial to the organization. Four tests for feasibility study are as follows :-

Technical Feasibility:

This is concerned with specifying equipment and software that will successfully satisfy the use considerably, but might include

- The feasibility to produce output in a given time because system is fast enough to handle multiple users.
- Response time under certain circumstances and ability to produce to process a certain volume of transaction.
- Feasibility to communicate data to distant location.

Economical Feasibility:

Economic analysis is the most frequent used technique used for evaluation the effectiveness of a proposed system. More commonly known as cost/benefit analysis the procedure is to determine the benefits and savings that are expected from a proposed system and compared them with cost. This System is Cost effective as there is no need of having separate peoples for evaluating the messages of peoples as System is smart enough to predict that.

Operational Feasibility:

It is mainly related to human organizational as social aspects. The points to be considered are – The system interface is standard, user friendly and provides extensive help. Hence no special training is required.

Social Feasibility:

Social feasibility is determination of whether a proposed project will be acceptable to people or not,So this project is Social and Feasible.

2.5 Objectives of Project

The main objective of this research is to produce a predictive model with better results that assess customer churn rate of companies using the predictive analytics algorithm for data mining. The supporting objectives examined are to:

I. Cluster customers into various categories to enhance marketing and promotional activities.
II. Mine the relevant patterns embedded in the collected data have a huge influence on the revenues and growth of the companies.

[CHAPTER-3]. Product Design

3.1 Product Perspective

Customer churn prediction is a classification technique. In this paper, we will explore the application of a customer dataset with a wide-variety of temporal features in order to create a highly-accurate customer churn model. In particular, we describe an effective method for handling temporally sensitive feature engineering.

3.2 Product Function

The main function of this model is:

1. Data Extraction from given 10 datasets
2. Data Requirement Gathering
3. Data Collection
4. Data Cleaning
5. Data Analysis
6. Data Interpretation
7. Data Visualization
8. Finding the accuracy

3.3 User Characteristics

Analysing user characteristics is an important aspect of any project.It allows us to clearly define and focus on who the end users are for the project. Also, it allows us to check the progress of the project to ensure that we are still developing the system for end users.

3.4 Constraints

- Churn prevention requires timely and accurate prediction of those customers who are likely to leave. Currently, such predictions are done by expert sales managers using limited number of variables and adhoc rules and algorithms on the available data.
    - Even when they are able to identify patterns within the data it is too late to take action and prevent customers from leaving.

- In reality, actual customer churn depends on numerous external and behavioral factors which themselves are not static and evolve in time. Hence it can be detected accurately only by taking into account all the variables as well as the changes in them over time.

3.5.1 Use Case Model/Flow Chart

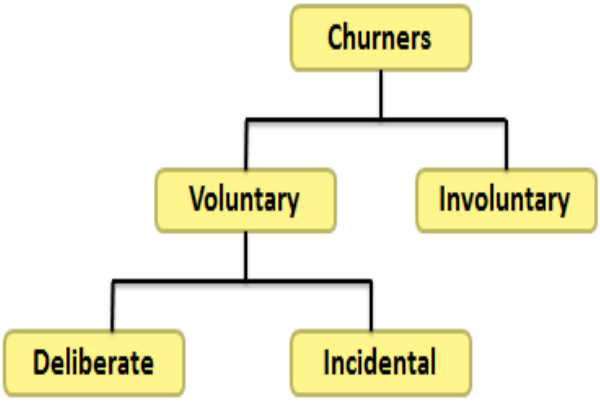


Fig.3.5(a)Type of churners

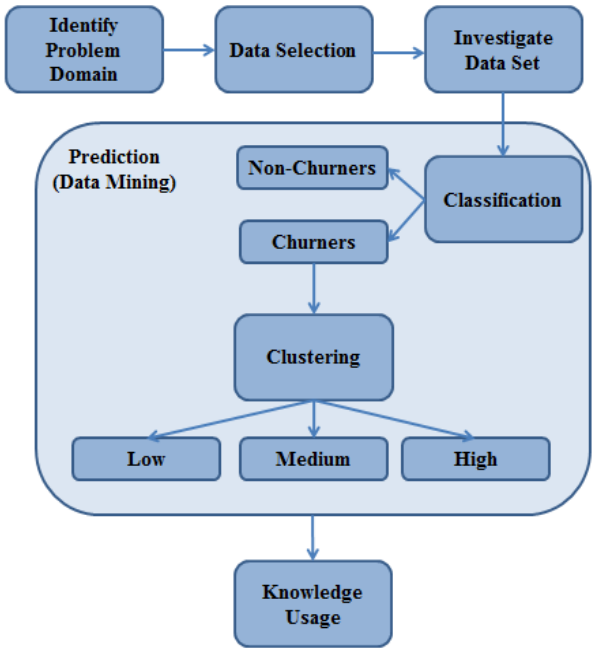


Fig 3.5(b)Flow of Whole model

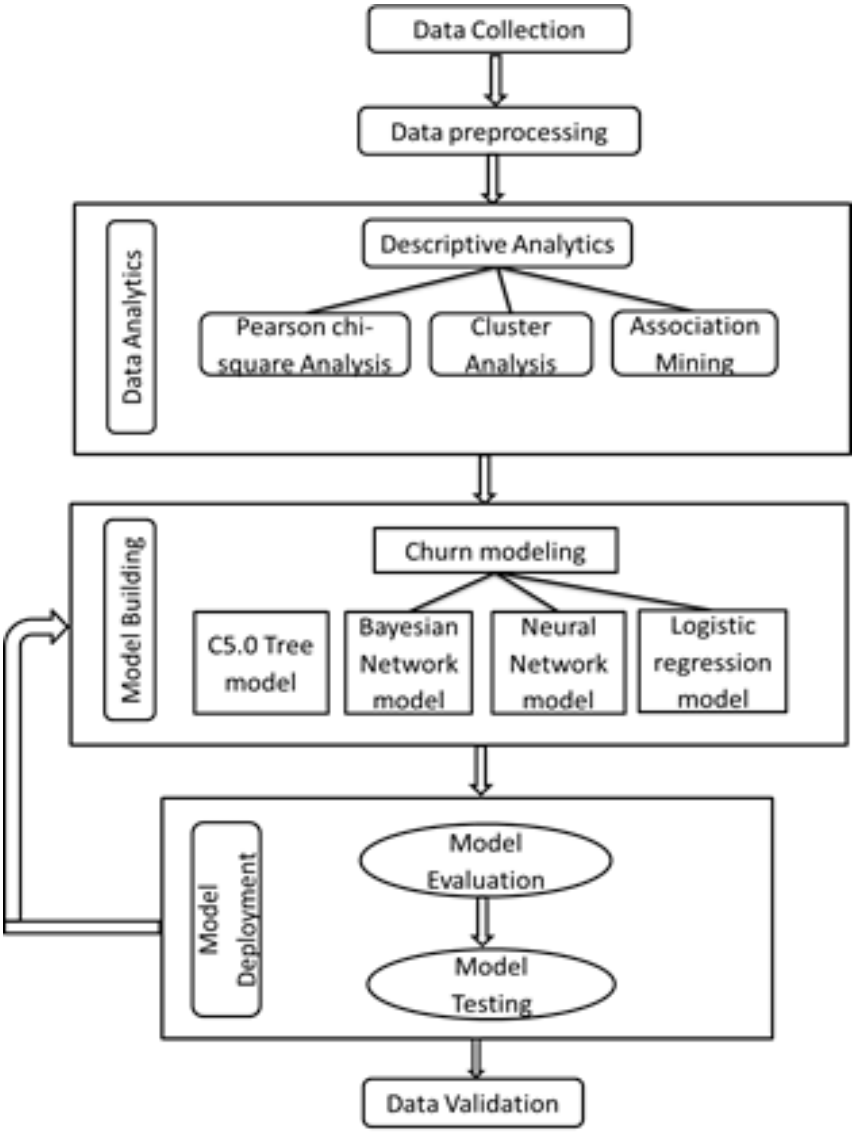


Fig.3.5(c) Step-By-Step process of Model

3.5.2 Gantt Chart

A Gantt chart is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph show the duration of each activity. Gantt charts illustrate the start and finish dates of the terminal elements and summary elements of a project**.**

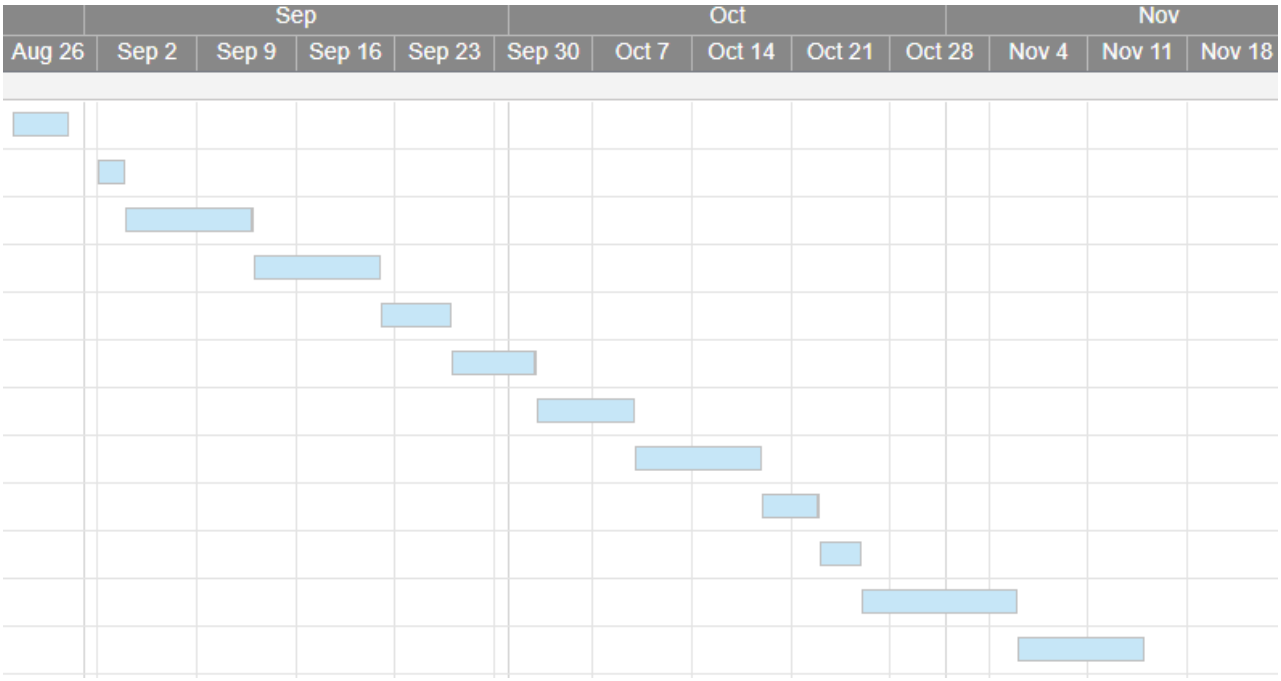


Fig.3.5.2(a) Gantt Chart

3.6 Assumptions and Dependencies

1. We assume that when the accuracy of the model falls below 90%, the prediction is not correct and we do some data treatment.
2. System speed is assumed to be fast enough for real time treatment of the datasets .
3. User is assumed to be familiar with the basic understanding and usage of Linux or any of the operating system.
4. Basic algorithm and data structure knowledge is required for the coding and implementation.

3.10 Specific Requirements

3.10.1 Software Requirements

- **Operating System:** Windows 7 or newer, 64-bit macOS 10.10+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.

- **Jupyter Notebook** : The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

3.10.2 Hardware Requirements

- System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 32-bit x86, 64-bit Power8/Power9.
- Minimum 5 GB disk space to download and install.

[Chapter-4] Development and Implementation

4.1 Introduction to Languages

Python:

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions- Python 2 and Python 3. Both are quite different. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. The software is developed using python and thus requires python to be installed on the user's computer. This applies to Linux users. Python(3.1) is used particularly used while testing with drone. Python is used mainly because of its huge set libraries which can be easily used for machine/deep learning tasks (for e.g. NumPy, Pytorch etc),One more reason for its usage is because of the huge community working on it makes it relatively much easier to find related articles of problems and prompt solutions. Also one more reason for its usage is that it has an easy implementation for OpenCV. What makes Python favourite for everyone is its powerful and easy implementation. Python is working at the backend mainly and is holding whole structure of project and neural network models are build using it in pysot which is a open source code of tracking based on neural networks. Python makes it lot easier to experiment with new ideas and code prototypes quickly in a language with minimal syntax like Python.

4.2 Any other Supporting Languages or tools

MS Excel:

Microsoft Excel is a spreadsheet program that is used to record and analyse numerical data. Think of a spreadsheet as a collection of columns and rows that form a table. Alphabetical letters are usually assigned to columns and numbers are usually assigned to rows. The point where a column and a row meet is called

a cell. The address of a cell is given by the letter representing the column and the number representing a row. Let's illustrate this using the following image.

PANDAS:

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.

In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data.

Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Key Features of Pandas

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations.
- High performance merging and joining of data.
- Time Series functionality.

Matplotlib:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc

Seaborn:

In the world of Analytics, the best way to get insights is by visualizing the data. Data can be visualized by representing it as plots which is easy to understand, explore and grasp. Such data helps in drawing the attention of key elements.

To analyse a set of data using Python, we make use of Matplotlib, a widely implemented 2D plotting library. Likewise, Seaborn is a visualization library in Python. It is built on top of Matplotlib.

It is summarized that if Matplotlib "tries to make easy things easy and hard things possible", Seaborn tries to make a well-defined set of hard things easy too."

Seaborn helps resolve the two major problems faced by Matplotlib; the problems are −

- Default Matplotlib parameters
- Working with data frames

As Seaborn compliments and extends Matplotlib, the learning curve is quite gradual. If you know Matplotlib, you are already half way through Seaborn.

Important Features of Seaborn:

Seaborn is built on top of Python's core visualization library Matplotlib. It is meant to serve as a complement, and not a replacement. However, Seaborn comes with some very important features. Let us see a few of them here. The features help in −

- Built in themes for styling matplotlib graphics
- Visualizing univariate and bivariate data
- Fitting in and visualizing linear regression models
- Plotting statistical time series data
- Seaborn works well with NumPy and Pandas data structures

4.3 Implementation of problem

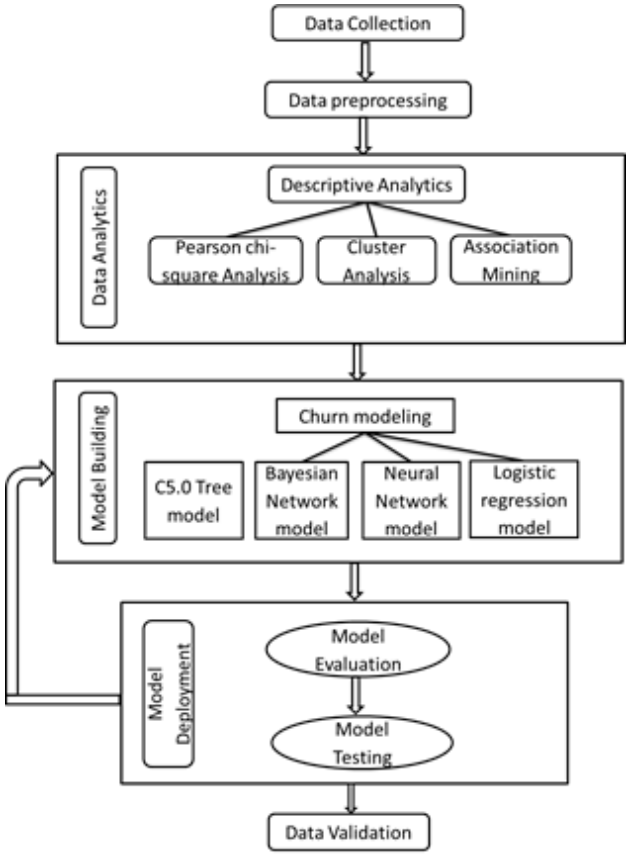


Fig 4.3(a) Working of the model

4.3.1 Data Extraction

Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern. Further data processing is done, which involves adding metadata and other data integration; another process in the data workflow.

The majority of data extraction comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.

Data in a warehouse may come from different sources, a data warehouse requires three different methods to utilize the incoming data. These processes are known as Extraction, Transformation, and Loading (ETL).

The process of data extraction involves retrieval of data from disheveled data sources. The data extracts are then loaded into the staging area of the relational database. Here extraction logic is used and source system is queried for data using application programming interfaces. Following this process, the data is now ready to go through the transformation phase of the ETL process.

4.3.2 Data Merging

Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary when you have raw data stored in multiple files, worksheets, or data tables, that you want to analyze all in one go.

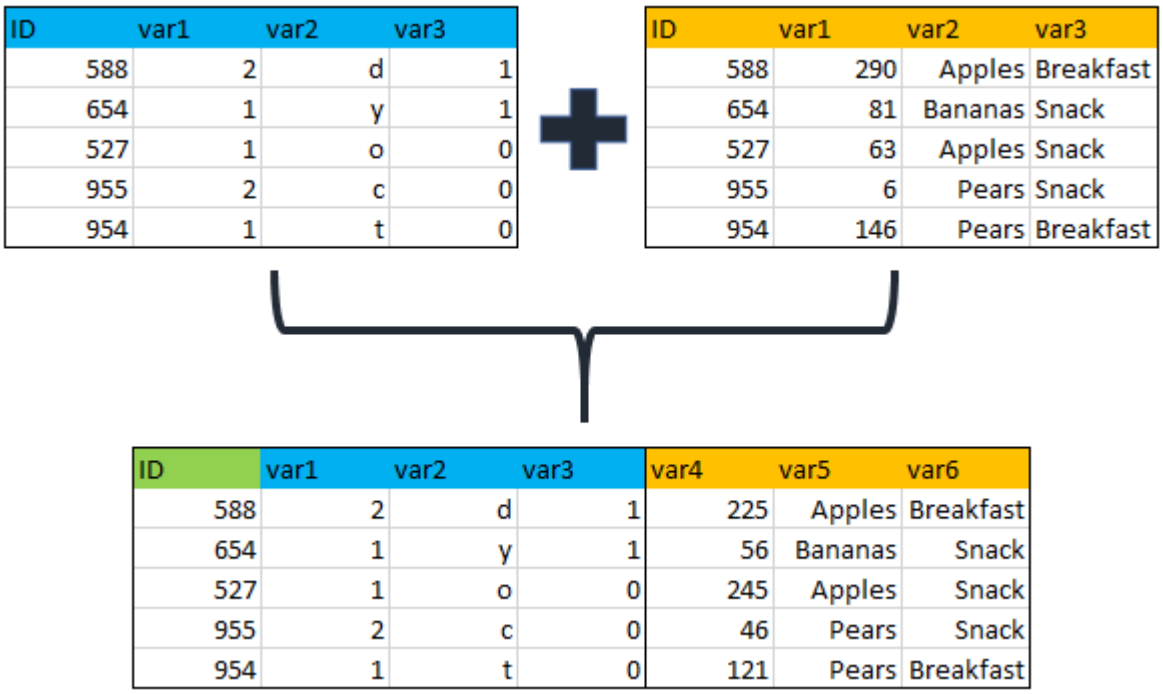
There are two common examples in which a data analyst will need to merge new cases into a main, or principal, data file:

They have collected data in a longitudinal study (tracker) – a projects in which an analyst collects data over a period of time and analyzes it as intervals.

They have collected data in a before-and-after project – where the analyst collects data before an event, and then again after.

Merging in New Variables:

Contrary to when you merge new cases, merging in new variables requires the IDs for each case in the two files to be the same, but the variable names should be different. In this scenario, which is sometimes referred to as augmenting your data (or in SQL, "joins") or merging data by columns (i.e. you're adding new columns of data to each row), you're adding in new variables with information for each existing case in your data file. As with merging new cases where not all variables are present, the same thing applies if you merge in new variables where some cases are missing – these should simply be given blank values.



ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0

ID	var1	var2	var3
588	290	Apples	Breakfast
654	81	Bananas	Snack
527	63	Apples	Snack
955	6	Pears	Snack
954	146	Pears	Breakfast

ID	var1	var2	var3	var4	var5	var6
588	2	d	1	225	Apples	Breakfast
654	1	y	1	56	Bananas	Snack
527	1	o	0	245	Apples	Snack
955	2	c	0	46	Pears	Snack
954	1	t	0	121	Pears	Breakfast

Fig.4.3.2(a) Merging

It could also happen that you have a new file with both new cases and new variables. The approach here will depend on the software you're using for your merge. If the software cannot handle merging both variables and cases at the same time, then consider first merging in only the new variables for the existing sample (i.e. augment first), and then append the new cases across all variables as a second step to your merge.

4.3.3 Data Pre-processing

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc.

In other words, whenever the **data** is gathered from different sources it is collected in raw format which is not feasible for the analysis.

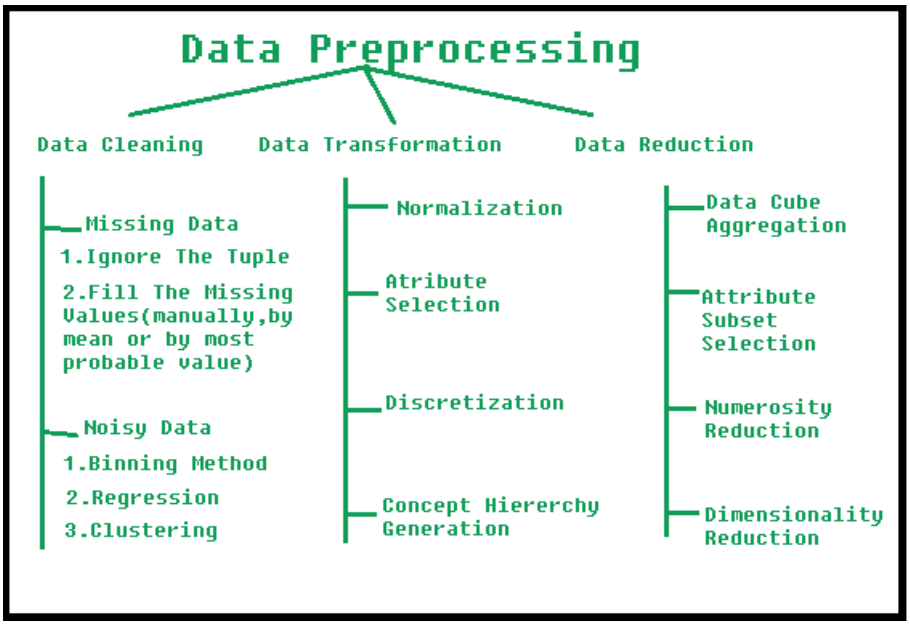


Fig.4.3.3(a) Steps involved in Data pre-processing

Steps Involved in Data Preprocessing**:**

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- Missing Data:

  This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

  1. Ignore the tuples:

     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

  2. Fill the Missing values:

     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- Noisy Data:

  Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

  1. Binning Method:

     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each

segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

2. Regression:
   Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
3. Clustering:
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

I. Normalization:
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
II. Attribute Selection:
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
III. Discretization:
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
IV. Concept Hierarchy Generation:
   Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

I. Data Cube Aggregation:
   Aggregation operation is applied to data for the construction of the data cube.
II. Attribute Subset Selection:
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.
III. Numerosity Reduction:
   This enable to store the model of data instead of whole data, for example: Regression Models.
IV. Dimensionality Reduction:
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Componenet Analysis).

4.3.4 Data Analytics

What is Data Analytics?

Data analytics is the science of analysing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system.

Why Data Analytics Matters?

Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.

A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

Types of Data Analytics:

Data analytics is broken down into four basic types.

1. Descriptive analytics describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
2. Diagnostic analytics focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
3. Predictive analytics moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
4. Prescriptive analytics suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.

Pairplot:

The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column. That creates plots as shown below.

The pairplot plot is shwon below. Its using the iris flower data set. The data set has 4 measurements: sepal width, sepal length, petal_length and petal_width. The data is mapped in the grid below. Because there are 4 measurements, it creates a 4x4 plot.Functions:

seaborn.pairplot

- Tidy (long-form) dataframe where each column is a variable and each row is an observation.
- Variable in data to map plot aspects to different colors.
- Order for the levels of the hue variable in the palette.
- Set of colors for mapping the hue variable.

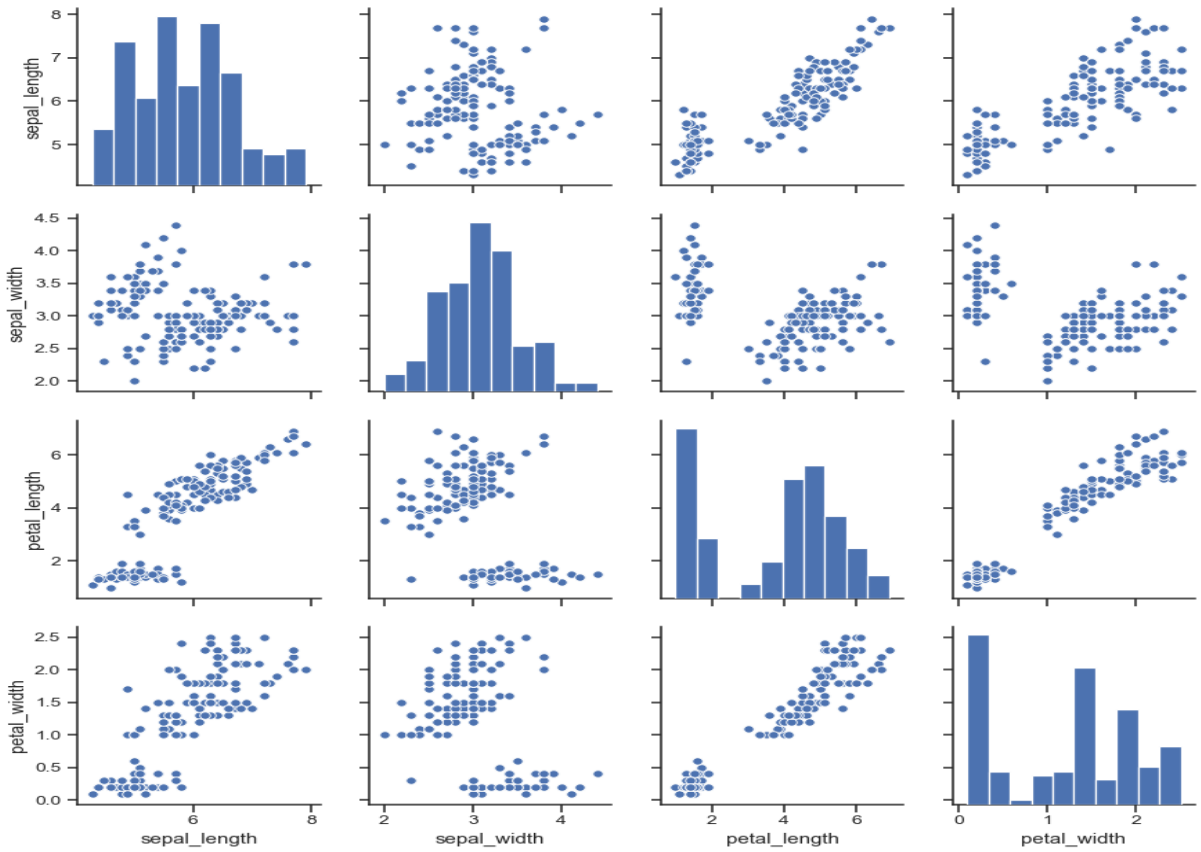


Fig.4.3.4(a) Pairplot

Countplot:

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot, so you can compare counts across nested variables.

seaborn.countplot(x=None, y=None, hue=None, data=None, order=None, hue_order=None, orient=None, color=None, palette=None, saturation=0.75, dodge=True, ax=None, **kwargs)
Show the counts of observations in each categorical bin using bars.

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot(), so you can compare counts across nested variables.

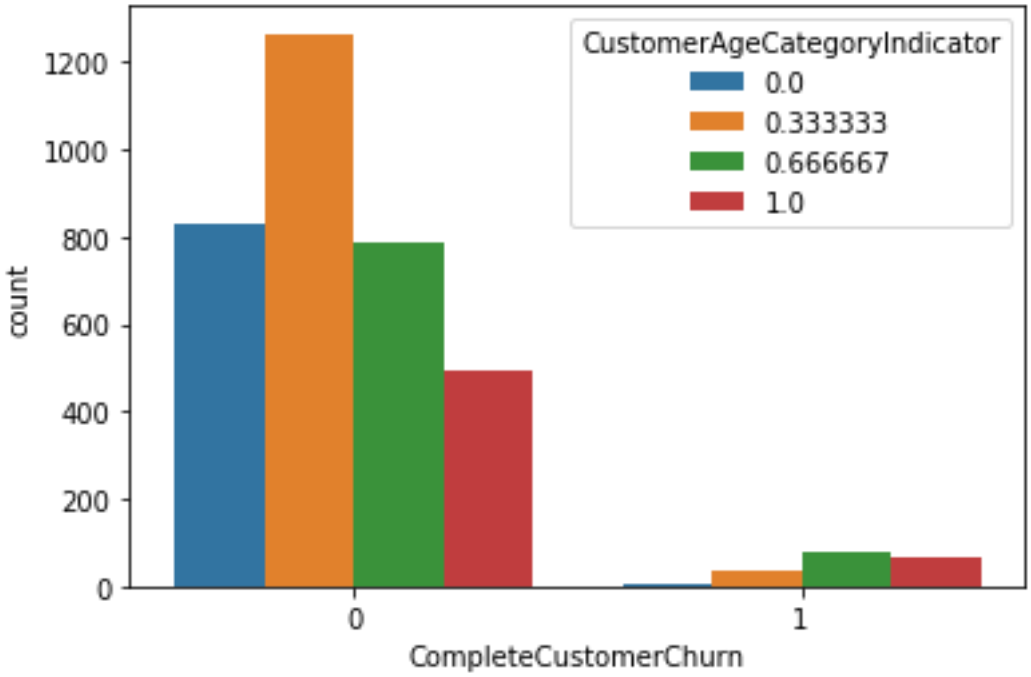


Fig.4.3.4(b) Countplot

Boxplot:

Boxplots are a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").

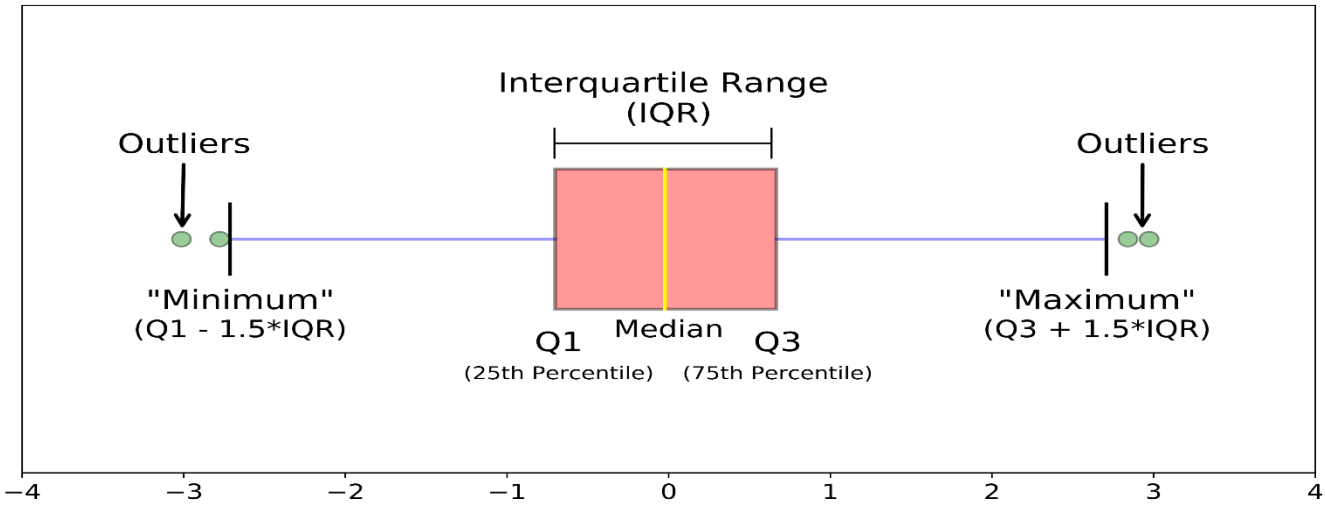


Fig.4.3.4(c) Boxplot

- Median (Q2/50th Percentile): the middle value of the dataset. First quartile (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the dataset.

- Third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the dataset.

- Interquartile range (IQR): 25th to the 75th percentile.

- Whiskers (shown in blue)

- Outliers (shown as green circles)

Heatmap:

A **heatmap** is a graphical representation of data that uses a system of color-coding to represent different values. **Heatmaps** are used in various forms of analytics but are most commonly used to show user behaviour on specific webpages or webpage templates

Fig.4.3.4(d) Heatmap Genarated

Piechart:

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

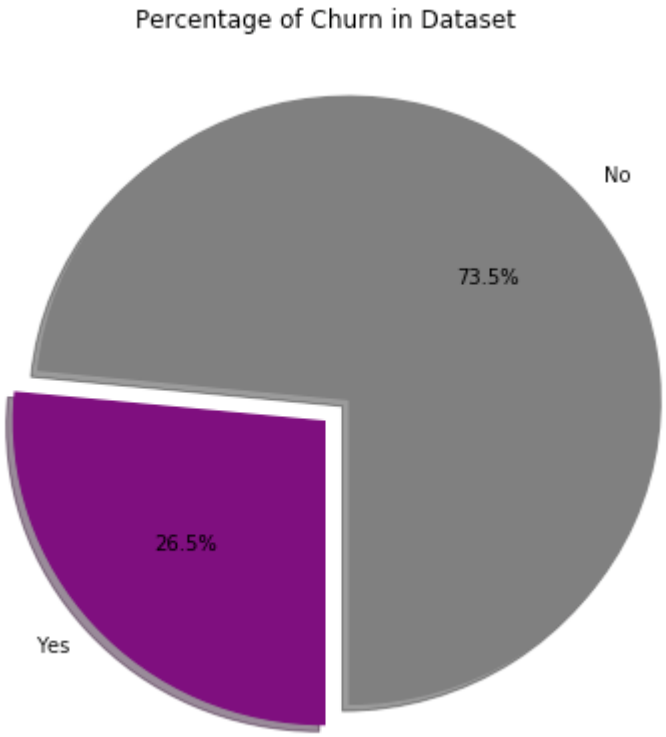


Fig.4.3.4(e) Pie-chart of the model

Demo of a basic pie chart plus a few additional features.

In addition to the basic pie chart, this demo shows a few optional features:

- slice labels
- auto-labeling the percentage
- offsetting a slice with "explode"
- drop-shadow
- custom start angle

4.3.5 Model Building

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value**.**

What are the types of logistic regression:

1. Binary (eg. Tumor Malignant or Benign)
2. Multi-linear functions failsClass (eg. Cats, dogs or Sheep's)

Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

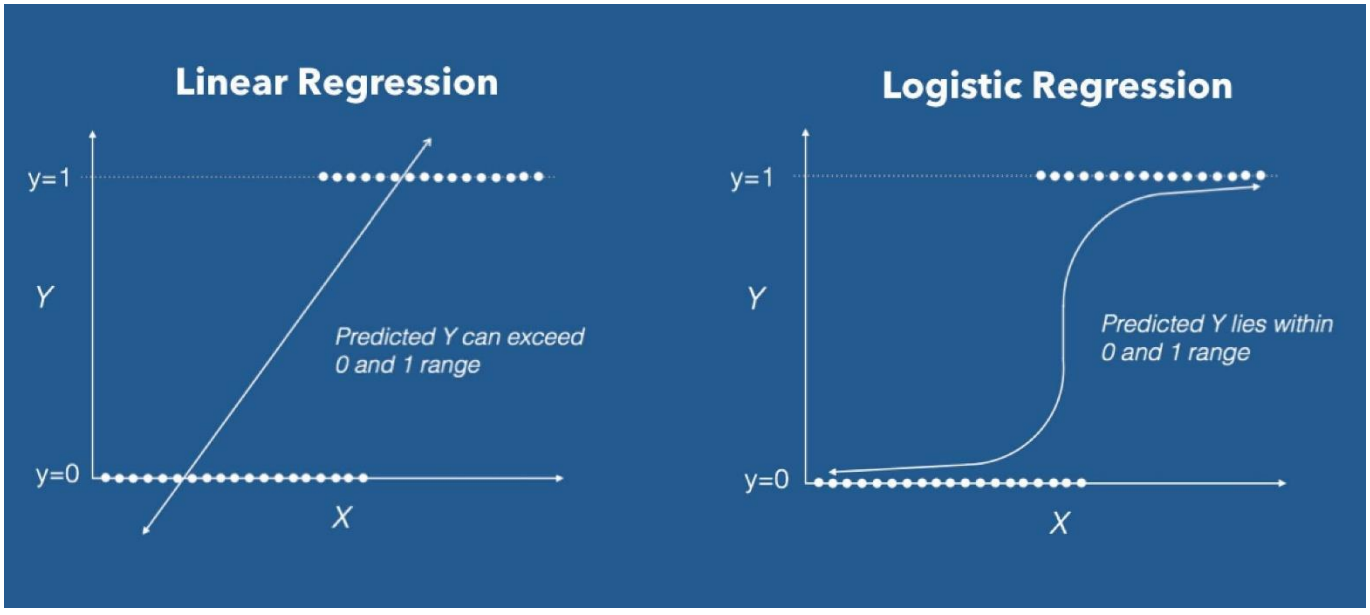


Fig.4.3.5(a) Linear and Logistic regression

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

What is the Sigmoid Function?

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

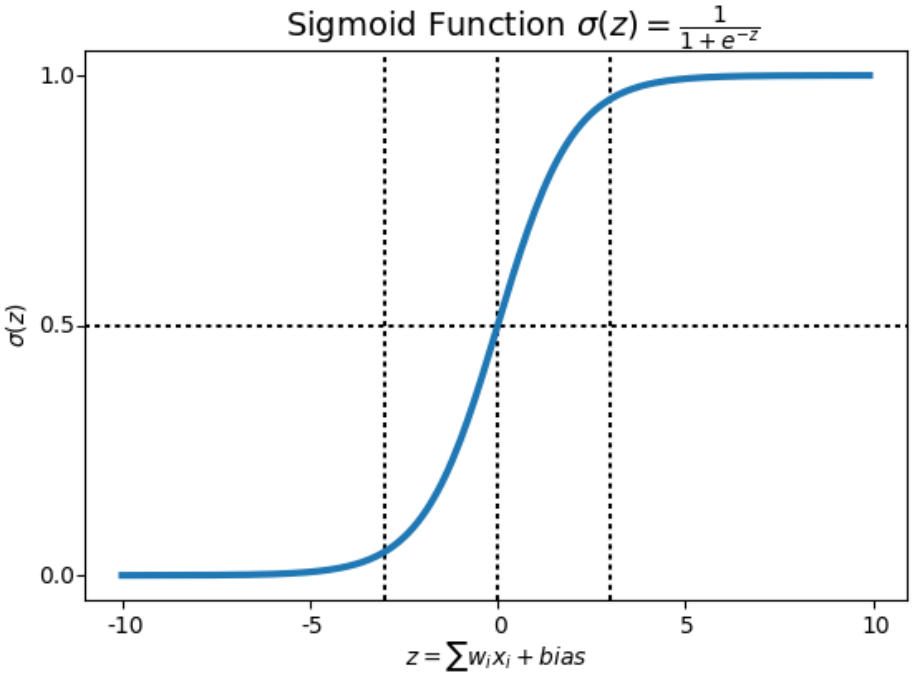


Fig.4.3.5(b) Graph of Sigmoid Function

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Hypothesis Representation

When using linear regression we used a formula of the hypothesis i.e.

$$h\Theta(x) = \beta_0 + \beta_1 X$$

For logistic regression we are going to modify it a little bit i.e.

$$\sigma(Z) = \sigma(\beta_0 + \beta_1 X)$$

We have expected that our hypothesis will give values between 0 and 1.

$$Z = \beta_0 + \beta_1 X$$

$h\Theta(x) = \text{sigmoid}(Z)$

i.e. $h\Theta(x) = 1/(1 + e^{-(\beta_0 + \beta_1 X)})$

Decision Boundary

We expect our classifier to give us a set of outputs or classes based on probability when we pass the inputs through a prediction function and returns a probability score between 0 and 1.

For Example, We have 2 classes, let's take them like cats and dogs(1 — dog , 0 — cats). We basically decide with a threshold value above which we classify values into Class 1 and of the value goes below the threshold then we classify it in Class 2.

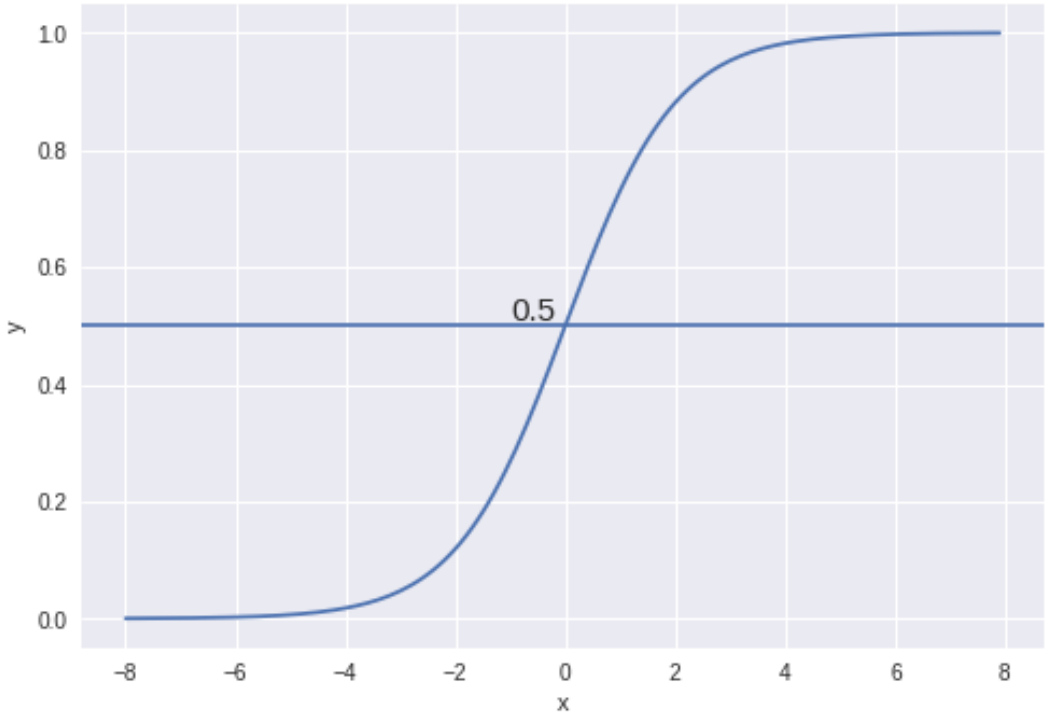


Fig.4.3.5(c) Decision boundry

Cost Function

We learnt about the cost function J(θ) in the Linear regression, the cost function represents optimization objective i.e. we create a cost function and minimize it so that we can develop an accurate model with minimum error.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2.$$

The Cost function of Linear regression

If we try to use the cost function of the linear regression in 'Logistic Regression' then it would be of no use as it would end up being a non-convex function with many local minimums, in which it would be very difficult to minimize the cost value and find the global minimum.

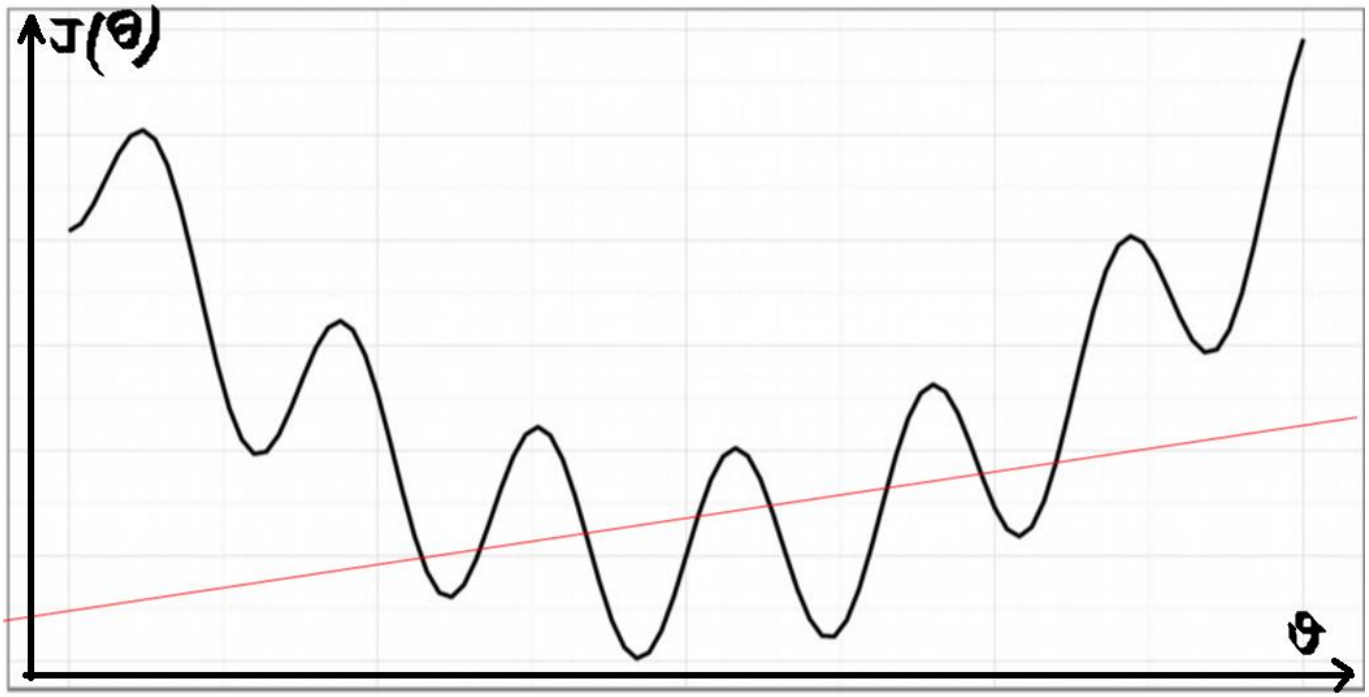


Fig.4.3.5(c) Fig representing the Cost function

4.4 Test cases

A TEST CASE is a set of actions executed to verify a particular feature or functionality of your software application. A Test Case contains test steps, test data, precondition, postcondition developed for specific test scenario to verify any requirement. The test case includes specific variables or conditions, using which a testing engineer can compare expected and actual results to determine whether a software product is functioning as per the requirements of the customer.

Our test case will going to depend on the p-value of all features. So for examining the **p-value**, we will print the summary table, to the p-values of all the features constituted with our model and datasets.

Below is the summary table with all the features after combining the 10 datasets:

Logit Regression Results

Dep. Variable:	CompleteCustomerChurn	No. Observations:	3558
Model:	Logit	Df Residuals:	3540
Method:	MLE	Df Model:	17
Date:	Fri, 13 Dec 2019	Pseudo R-squ.:	0.5636
Time:	21:24:55	Log-Likelihood:	-319.80
converged:	False	LL-Null:	-732.87
Covariance Type:	nonrobust	LLR p-value:	1.224e-164

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	9.8684	1.444	6.833	0.000	7.038	12.699
CustomerAgeCategoryIndicator	2.9257	0.318	9.195	0.000	2.302	3.549
AcctPaymentStatusIndicator	-2.5521	0.989	-2.581	0.010	-4.490	-0.614
RepChangesinLast002Months_SUM	0.9354	0.622	1.505	0.132	-0.283	2.154
NumberOfProductionInstances_SUM	-114.9742	1.45e+04	-0.008	0.994	-2.85e+04	2.82e+04
OptyWonLast12Months_Percent_SUM	-2.6113	0.289	-9.041	0.000	-3.177	-2.045
OverallEngagementSatisfactionIndicator_SUM	-2.2551	1.228	-1.837	0.066	-4.662	0.151
FulfillerPriceCategoryIndicator	-0.3410	0.323	-1.055	0.291	-0.974	0.292
KnowledgeEventAttendeeCount	-11.1694	4.080	-2.738	0.006	-19.165	-3.174
TotalAdoptionPercent	-71.3452	7.955	-8.969	0.000	-86.936	-55.754
LicensedAdoptionPercent	1.2419	1.053	1.179	0.238	-0.823	3.306
AdoptionCategoryIndicator	29.4649	3.301	8.927	0.000	22.995	35.934
LicensingCategoryIndicator	0.6048	0.092	6.542	0.000	0.424	0.786
ResponseRatePercent	-1.4892	1.044	-1.426	0.154	-3.536	0.558
NPSSatisfactionIndicator_SUM	-10.9554	2.859	-3.832	0.000	-16.558	-5.352
OverallSupportSurveySatisfactionIndicator_SUM	-6.8932	2.086	-3.304	0.001	-10.982	-2.804
UsageRiskIndicator_SUM	-1.0000	0.763	-1.310	0.190	-2.496	0.496
TotalActivityCount_SUM	-0.3984	0.442	-0.901	0.368	-1.265	0.468

Possibly complete quasi-separation: A fraction 0.39 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

What is p-value?

When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.

The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population). The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- p-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

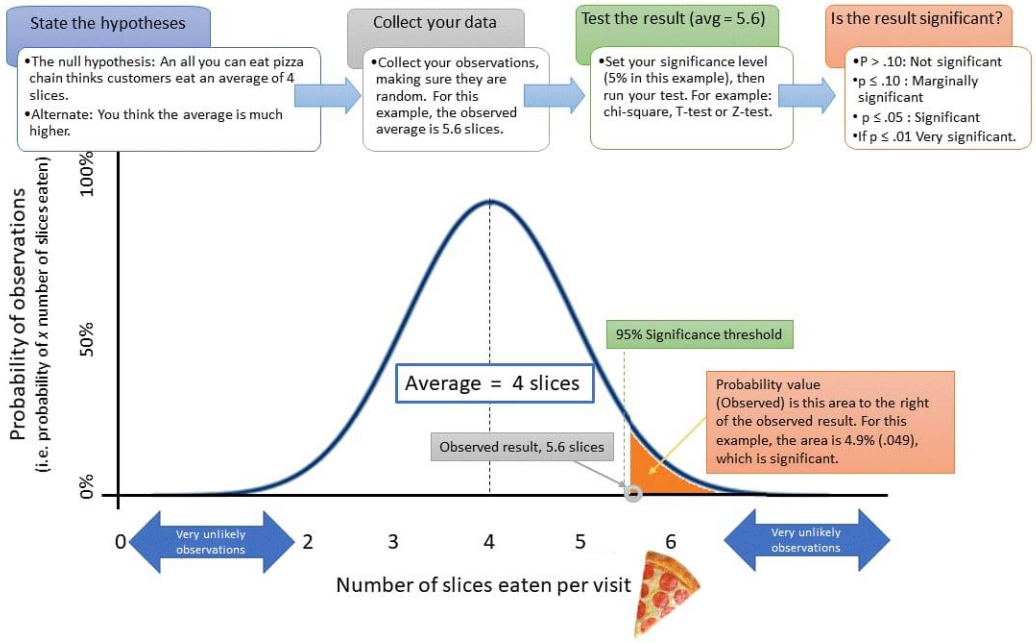


Fig.4.4(a) p-value explained

So, now on the basis of p-value, we will get our final test case by removing features with more p-value. And finally we get our all necessary required features for our model with highest accuracy.

Logit Regression Results

Dep. Variable:	CompleteCustomerChurn	No. Observations:	3558
Model:	Logit	Df Residuals:	3547
Method:	MLE	Df Model:	10
Date:	Fri, 13 Dec 2019	Pseudo R-squ.:	0.5532
Time:	21:25:00	Log-Likelihood:	-327.44
converged:	True	LL-Null:	-732.87
Covariance Type:	nonrobust	LLR p-value:	9.550e-168

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	9.7983	1.352	7.245	0.000	7.147	12.449
CustomerAgeCategoryIndicator	2.8621	0.314	9.126	0.000	2.247	3.477
AcctPaymentStatusIndicator	-2.5417	0.914	-2.781	0.005	-4.333	-0.751
OptyWonLast12Months_Percent_SUM	-2.6112	0.285	-9.170	0.000	-3.169	-2.053
OverallEngagementSatisfactionIndicator_SUM	-2.2301	1.219	-1.829	0.067	-4.620	0.159
KnowledgeEventAttendeeCount	-11.9750	3.946	-3.035	0.002	-19.708	-4.242
TotalAdoptionPercent	-74.7488	7.940	-9.414	0.000	-90.311	-59.186
AdoptionCategoryIndicator	31.0146	3.292	9.421	0.000	24.562	37.467
LicensingCategoryIndicator	0.5839	0.085	6.848	0.000	0.417	0.751
NPSSatisfactionIndicator_SUM	-12.4009	2.782	-4.457	0.000	-17.854	-6.948
OverallSupportSurveySatisfactionIndicator_SUM	-6.7958	1.992	-3.412	0.001	-10.699	-2.893

[Chapter -5] Conclusion and Future Scope

5.1 Conclusion

Churn prediction is a function that involves systematic analysis of customer data for identifying and analyzing patterns and trends of customer loyalty and blend. The detected patterns and trends can be used by telecommunication industries to improve customer relationship and at the same time improve net profit. Identification of churners and nonchurners is a time consuming and critical task, that has to be performed carefully, as the future growth of the company relies on the result of such an analysis. This task is considered challenging because of two reasons, (i) customer information volume has increased and (ii) the data available is inconsistent and are incomplete thus making the task of formal analysis a difficult task. Further, due to its vast size, investigation and analysis of customer database takes longer duration due to the complexity of these issues.

The current needs of telecom companies is a tool that can be used to help them to understand customer patterns and locate churners and possible actions that can be taken to convert the churners to non-churners. This tool is called as ‘Customer Loyalty Assessment Model and Actionable Knowledge Discovery System’ and the main goal is to provide timely and pertinent customer information to decision-makers in a company. The present research work focus on developing such a system that can be used by telecom industry easily discover customer patterns and trends, make forecasts, find relationships and possible explanations and identify possible churners. The proposed system proposes the use of data mining techniques during the design and development

5.2 Future Scope

Future research in this area should be around identifying the methods used by vendors for their analytical tools and identifying the strengths and weaknesses of these techniques. In terms of churn analysis, it would also be interesting to discover which customer variables are used by the companies that do provide customer churn analysis, the estimated accuracy rates for these predictions and if improvements on the accuracy could be made by using alternative technologies and added variables. It is anticipated that the information needed for this proposal of future research would be very difficult to acquire and would most certainly require conversations with developers from the vendors in question.

References

1) Boris Babenko, Ming-Hsuan Yang, Serge Belongie. Visual Tracking with Online Multiple Instance Learning, Honda Research Institute, USA, 2009.

2) Jianjun Ni,Xue Zhang,Pengfei Shi,Jinxiu Zhu. An Improved Kernelized Correlation Filter Based Visual Tracking Method,Hohai University, Changzhou 213022, China, 2018.

3) Zdenek Kalal, Krystian Mikolajczyk, Jiri Matas. Tracking-Learning-Detection, IEEE Transactions on pattern analysis and machine intelligence, Vol. 6, No. 1, 2010.

4) Gregory Koch, Richard Zemel, Ruslan Salakhutdinov. Siamese Neural Networks for One-shot Image Recognition, Department of Computer Science, University of Toronto. Toronto, Ontario, Canada, 2016.

5) Jack Valmadre, Luca Bertinetto, João F. Henriques, Andrea Vedaldi, Philip H.S. Torr. End-to-end representation learning for Correlation Filter based tracking, University of Oxford, CVPR, 2017.

6) Qing Guo, Wei Feng , Ce Zhou , Rui Huang, Liang Wan, Song Wang. Learning Dynamic Siamese Network for Visual Object Tracking, IEEE Xplore, International Conference on Computer Vision, 2017.

7) Bo Li , Junjie Yan,Wei Wu, Zheng Zhu, Xiaolin Hu. High Performance Visual Tracking with Siamese Region Proposal Network. IEEE Xplore, CVPR, 2018.

8) Zheng Zhu, Qiang Wang, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware Siamese Networks for Visual Object Tracking. European Conference on Computer Vision, 2018.

9) Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, Junjie Yan. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. arXiv:1812.11703, 2018.

10) Fangyi Zhang, Qiang Wang, Bo Li. SenseTime Research platform for single object tracking, implementing algorithms like SiamRPN and SiamMask (https://github.com/STVIR/pysot), 2019.

11) ROS Documentation (http://wiki.ros.org/Documentation).