A PROJECT ON HEALTH INSURANCE LEAD PREDICTION USING PYTHON

Project submitted and prepared by Bishal Biswas
under guidance of Prof. Sambita Chakraborty

Project done as a part of assignments for
Post Graduate Diploma Program
From
Bombay Stock Exchange (BSE)
In collaboration with
Maulana Abul Kalam Azad University of Technology
(MAKAUT)

Certification of Approval

This document is hereby approved as credible study of the science subject carried out and represented in a manner to satisfy to the warrants of its acceptance as a prerequisite to the degree for which it has been submitted.

Moreover, it is understood that by this approval the undersigned does not necessarily endorse or approve any statements made, the opinion expressed or conclusion drawn therein but approved only for the sole purpose for which it has been indeed submitted.

Signatures of the Examiners with date.

×________________________________

×________________________________

×________________________________

Dated:

Countersigned by:

×________________________________

Prof. Sambita Chakraborty

Acknowledgement

Our project and everything started during the ending rule of SARS – CoVID19, virtually crippling the society and world as a whole sending everything into a lockdown, although at better stage but still this course of Post Graduate Diploma in Data Science by BSE in collaboration with MAKAUT was made possible thanks to the diplomacy and steps taken by both institutes to combat the situation and make this course and project a possibility.

I want to take this opportunity of the project to thank the people at BSE and MAKAUT who provided us this opportunity to have an exposure to real life scenarios and the status of the present market. I also want to thank Prof. Sambita Chakraborty for guiding with every step from imparting knowledge about the subject to the intricacies of the Data Preparation and Analysis including Cleaning, Visualization, clearing doubts and issues faced in addition to solving problems encountered.

I am also grateful to my batchmates and peers where our collective knowledgebase and doubt clearing helped a lot in completing this project. Lastly, I want to thank my family for the mental support they provided me that played a big part in completing this project.

×______________________________

Bishal Biswas.
BSE GENERATED ID: PGDDSPJULY2020/1
MAKAUT ENROLMENT:20BIL001P12029005
MAKAUT APPLICATION ID: 91268
b.biswas_94587@ieee.org

Contents

Objective and Purpose

Data is truly considered a resource in today's world. Data is everywhere and part of our daily lives in more ways than most of us realize in our daily lives. The amount of digital data that exists—that we create—is growing exponentially. As per the World Economic Forum, by 2025 we will be generating about 463 exabytes of data globally per day. Hence, there is a need for professionals who understand the basics of data science, big data, and data analytics. These three terms are often heard frequently in the industry, and while their meanings share some similarities, they also mean different things.

Data science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning data. This umbrella term includes various techniques that are used when extracting insights and information from data.

Now, Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. A buzzword that is used to describe immense volumes of data, both unstructured and structured, big data can inundate a business on a day-to-day basis. Big data is used to analyze insights, which can lead to better decisions and strategic business moves.

Gartner provides the following definition of big data: "Big data is high-volume, and high-velocity or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

Data analytics involves applying an algorithmic or mechanical process to derive insights and running through several data sets to look for meaningful correlations. It is used in several industries, which enables organizations and data analytics companies to make more informed decisions, as well as verify and disprove existing theories or models. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

Industries like IT, Retail, Manufacturing, Automobile, Financial Institute, E Commerce etc. are focusing in depth towards Big Data Concept because they have found out its importance, they know Data is Asset and its value will grow day by day and it can lead the Global business. Some benefits of it are:

- Data driven decision making with more accuracy.
- Customer active engagement.
- Operation optimization.
- Data driven Promotions.

- Preventing frauds & threats.
- Exploring new sources of revenue.
- Being ahead of your competitors.

Thus, the growth of big data analytics will also probably be good for data scientists, especially those who have strong backgrounds in big data. Based on the growth of the big data analytics market in the past few years, along with the rising number of job openings, it's likely that demand for these skills will continue to increase in the near future.

Introduction

Since the invention of computers, people have used the term data to refer to computer information, and this information was either transmitted or stored. But that is not the only data definition; there exist other types of data as well. So, what is the data? Data can be texts or numbers written on papers, or it can be bytes and bits inside the memory of electronic devices, or it could be facts that are stored inside a person's mind.

Now, if we talk about data mainly in the field of science, then the answer to "what is data" will be that data is different types of information that usually is formatted in a particular manner. All the software is divided into two major categories, and those are programs and data. Programs are the collection made of instructions that are used to manipulate data. So, now after thoroughly understanding what is data and data science, let us learn some fantastic facts.

Growth in the field of technology, specifically in smartphones has led to text, video, and audio is included under data plus the web and log activity records as well. Most of this data is unstructured.

The term Big Data is used in the data definition to describe the data that is in the petabyte range or higher. Big Data is also described as 5Vs: variety, volume, value, veracity, and velocity. Nowadays, web-based eCommerce has spread vastly, business models based on Big Data have evolved, and they treat data as an asset itself. And there are many benefits of Big Data as well, such as reduced costs, enhanced efficiency, enhanced sales, etc.

The meaning of data expands beyond the processing of data in computing applications. When it comes to what data science is, a body made of facts is called data science. Accordingly, finance, demographics, health, and marketing also have different meanings of data, which ultimately make up different answers for what is data.

When we talk about data, we usually think of some large datasets with huge number of rows and columns. While that is a likely scenario, it is not always the case — data could be in so many different forms: Structured Tables, Images, Audio files, Videos etc. Machines don't understand free text, image or video data as it is, they understand 1s and 0s. So, it probably won't be good enough if we put on a slideshow of all our images and expect our machine learning model to get trained just by that, hence a need for processing the data is required before being fed to the system for any analysis or estimation making learning or prediction too farfetched an idea to see realization.

Clean data is crucial for insightful data analysis. Data cleansing, data cleaning or data scrubbing is the first step in the overall data preparation process. It is the process of analyzing, identifying and correcting messy, raw data. Data cleaning involves filling in missing values, identifying and fixing errors and determining if all the information is in the right rows and columns. When analyzing organizational data to make strategic decisions you must start with a thorough data cleansing process. Cleaning data is crucial to data analysis. Data cleaning lays the groundwork for efficient, accurate and effective data analysis. Without cleaning data beforehand, the analysis process won't be clear or as accurate because the information in the dataset will be unorganized and scattered. Good analysis rests on clean data–it's as simple as that.

Analysis is the process of breaking a complex topic or substance into smaller parts in order to gain a better understanding of it. The technique has been applied in the study of mathematics and logic since before Aristotle, though analysis as a formal concept is a relatively recent development. Implementing these ideas of analysis using statistical processes to determine the future or optimize situations using available data at the disposal or data obtained from a specific source is the very idea on which the Data Analysis and subsequently Big Data Analysis is based on.

Beyond the analysis, even if we stop at this point, we are just holding a set of numbers, some arranged others not but all equivalently shared between fields of significance, which if not properly visualized wouldn't make any sense to a person working beyond the realm or field from where the data is based. Hence, the need for visualization stands. Visualization is the process of putting together visual mental imagery of what you are wanting to manifest. Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data Gathering

Data gathering or Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.

Data collection can be carried by from various ways like:

1. Surveys. Surveys are one way in which you can directly ask customers for information.
2. Online Tracking.
3. Transactional Data Tracking.
4. Online Marketing Analytics.
5. Social Media Monitoring.
6. Collecting Subscription and Registration Data.
7. In-Store Traffic Monitoring.

Primary Data Collection

The term "primary data" refers to data you collect yourself, rather than data you gather after another party initially recorded it. Primary data is information obtained directly from the source. You will be the first party to use this exact set of data.

When it comes to data businesses collect about their customers, primary data is also typically first-party data. First-party data is the information you gather directly from your audience. It could include data you gathered from online properties, data in your customer relationship management system or non-online data you collect from your customers through surveys and various other sources.

First-party data differs from second-party and third-party data. Second-party data is the first-party data of another company. You can purchase second-party data directly from the organization that collected it or buy it in a private marketplace. Third-party data is information a company has pulled together from numerous sources. You can buy and sell this kind of data on a data exchange, and it typically contains a large number of data points. Because first-party data comes directly from your audience, you can have high confidence in its accuracy, as well as its relevance to your business.

Second-party data has many of the same positive attributes as first-party data. It comes directly from the source, so you can be confident in its accuracy, but it also gives you insights you couldn't get with your first-party data. Third-party data offers much more scale than any other type of data, which is its primary benefit.

Different types of data can be useful in different scenarios. It can also be helpful to use different types of data together. First-party data will typically be the foundation of your dataset. If your first-party data is limited, though, you may want to supplement it with second-party or third-party data. Adding these other types of data can increase the scale of your audience or help you reach new audiences.

Quantitative vs. Qualitative Data

Quantitative data comes in the form of numbers, quantities and values. It describes things in concrete and easily measurable terms. Examples include the number of customers who bought a given product, the rating a customer gave a product out of five stars and the amount of time a visitor spent on your website.

Because quantitative data is numeric and measurable, it lends itself well to analytics. When you analyze quantitative data, you may uncover insights that can help you better understand your audience. Because this kind of data deals with numbers, it is very objective and has a reputation for reliability.

Qualitative data is descriptive, rather than numeric. It is less concrete and less easily measurable than quantitative data. This data may contain descriptive phrases and opinions. Examples include an online review a customer writes about a product, an answer to an open-ended survey question about what type of videos a customer likes to watch online and the conversation a customer had with a customer service representative.

Qualitative data helps explains the “why” behind the information quantitative data reveals. For this reason, it is useful for supplementing quantitative data, which will form the foundation of your data strategy. Because quantitative data is so foundational, this article will focus on collection methods for quantitative primary data.

Data Preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data. Good data preparation allows for efficient analysis, limits errors and inaccuracies that can occur to data during processing, and makes all processed data more accessible to users. It’s also gotten easier with new tools that enable any user to cleanse and qualify data on their own.

Data preparation is often a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.

Benefits of Data Preparation

76% of data scientists say that data preparation is the worst part of their job, but the efficient, accurate business decisions can only be made with clean data. Data preparation helps:

- Fix errors quickly — Data preparation helps catch errors before processing. After data has been removed from its original source, these errors become more difficult to understand and correct.

- Produce top-quality data — Cleaning and reformatting datasets ensures that all data used in analysis will be high quality.
- Make better business decisions — Higher quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient and high-quality business decisions.

Integrating the Cloud technology to the data preparation

As data and data processes move to the cloud, data preparation moves with it for even greater benefits, such as:

- Superior scalability — Cloud data preparation can grow at the pace of the business. Enterprise don't have to worry about the underlying infrastructure or try to anticipate their evolutions.
- Future proof — Cloud data preparation upgrades automatically so that new capabilities or problem fixes can be turned on as soon as they are released. This allows organizations to stay ahead of the innovation curve without delays and added costs.
- Accelerated data usage and collaboration — Doing data prep in the cloud means it is always on, doesn't require any technical installation, and lets teams collaborate on the work for faster results.

Considering beyond the above points, a good, cloud-native data preparation tool will offer other benefits (like an intuitive and simple to use GUI) for easier and more efficient preparation.

Data Preparation Steps

The specifics of the data preparation process vary by industry, organization and need, but the framework remains largely the same.

1. Gather data: The data preparation process begins with finding the right data. This can come from an existing data catalog or can be added ad-hoc.
2. Discover and assess data: After collecting the data, it is important to discover each dataset. This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context.
3. Cleanse and validate data: Cleaning up the data is traditionally the most time-consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps. Important tasks here include:
    a. Removing extraneous data and outliers.
    b. Filling in missing values.
    c. Conforming data to a standardized pattern.
    d. Masking private or sensitive data entries.

   Once data has been cleansed, it must be validated by testing for errors in the data preparation process up to this point. Often times, an error in the system will become apparent during this step and will need to be resolved before moving forward.

4. Transform and enrich data: Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience. Enriching data refers to adding and connecting data with other related information to provide deeper insights.
5. Store data: Once prepared, the data can be stored or channeled into a third-party application—such as a business intelligence tool—clearing the way for processing and analysis to take place.

Data Cleaning

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results. There are several methods for cleaning data depending on how it is stored along with the answers being sought.

Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set's accuracy without necessarily deleting information. For one, data cleaning includes more actions than removing data, such as fixing spelling and syntax errors, standardizing data sets, and correcting mistakes such as empty fields, missing codes, and identifying duplicate data points. Data cleaning is considered a foundational element of the data science basics, as it plays an important role in the analytical process and uncovering reliable answers. Most importantly, the goal of data cleaning is to create data sets that are standardized and uniform to allow business intelligence and data analytics tools to easily access and find the right data for each query.

Regardless of the type of analysis or data visualizations you need, data cleaning is a vital step to ensure that the answers you generate are accurate. When collecting data from several streams and with manual input from users, information can carry mistakes, be incorrectly inputted, or have gaps. Data cleaning helps ensure that information always matches the correct fields while making it easier for business intelligence tools to interact with data sets to find information more efficiently. One of the most common data cleaning examples is its application in data warehouses.

A successful data warehouse stores a variety of data from disparate sources and optimizes it for analysis before any modeling is done. To do so, warehouse applications must parse through millions of incoming data points to make sure they're accurate before they can be slotted into the right database, table, or other structure. Organizations that collect data directly from consumers filling in surveys, questionnaires, and forms also use data cleaning extensively. In their cases, this includes checking that data was entered in the correct field, that it doesn't feature invalid characters, and that there are no gaps in the information provided.

Data Analysis

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analyzing our past or future and making decisions based on it. For that, we gather memories of our past or dreams of our future. So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.

Types of Data Analysis: Techniques and Methods

There are several types of Data Analysis techniques that exist based on business and technology. However, the major Data Analysis methods are:

- Text Analysis
- Statistical Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

Text Analysis

Text Analysis is also referred to as Data Mining. It is one of the methods of data analysis to discover a pattern in large data sets using databases or data mining tools. It used to transform raw data into business information. Business Intelligence tools are present in the market which is used to take strategic business decisions. Overall, it offers a way to extract and examine data and deriving patterns and finally interpretation of the data.

Statistical Analysis

Statistical Analysis shows "What happen?" by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modeling of data. It analyses a set of data or a sample of data.

There are two categories of this type of Analysis - Descriptive Analysis and Inferential Analysis.

- Descriptive Analysis: analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.
- Inferential Analysis: analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples.

Diagnostic Analysis

Diagnostic Analysis shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data. If a new problem arrives in your business process, then you can look into this Analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

Predictive Analysis

Predictive Analysis shows "what is likely to happen" by using previous data. The simplest data analysis example is like if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses. But of course, it's not easy like this because you have to think about other circumstances like chances of prices of clothes is increased this year or maybe instead of dresses you want to buy a new bike, or you need to buy a house!

So here, this Analysis makes predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig in it.

Prescriptive Analysis

Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

Exploratory Data Analysis ~ EDA

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;

5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

History of EDA

EDA holds its roots from the seminal work in EDA that is Exploratory Data Analysis, Tukey, (1977). Over the years it has benefitted from other noteworthy publications such as Data Analysis and Regression, Mosteller and Tukey (1977), Interactive Data Analysis, Hoaglin (1977), The ABC's of EDA, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.

Techniques of EDA

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (such as data traces, histograms, bi-histograms, probability plots, lag plots, block plots, and Youden plots.
2. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

Exploratory Data Analysis vs Classical Data Analysis (EDA vs CDA)

EDA is a data analysis approach. Besides EDA other data analysis approaches also exist and now question arises how does EDA differ from these other approaches. Three popular data analysis approaches are:

1. Classical
2. Exploratory (EDA)
3. Bayesian

These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

- For classical analysis, the sequence is

  Problem → Data → Model → Analysis → Conclusions

- For EDA, the sequence is

  Problem → Data → Analysis → Model → Conclusions

- For Bayesian, the sequence is

  Problem → Data → Model → Prior Distribution → Analysis → Conclusions

Thus, for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model. For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate. Finally, for a Bayesian analysis, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

In the real world, data analysts freely mix elements of all of the above three approaches (and other approaches). The above distinctions were made to emphasize the major differences among the three approaches.

Focusing on EDA versus classical, these two approaches differ as follows:

1. Models
2. Focus
3. Techniques
4. Rigor

5. Data Treatment
6. Assumptions

Data Interpretation

Data interpretation is the process of reviewing data through some predefined processes which will help assign some meaning to the data and arrive at a relevant conclusion. It involves taking the result of data analysis, making inferences on the relations studied, and using them to conclude.

Therefore, before one can talk about interpreting data, they need to be analyzed first. From the previous two sections we know that data analysis is the process of ordering, categorizing, manipulating, and summarizing data to obtain answers to research questions. It is usually the first step taken towards data interpretation. It is evident that the interpretation of data is very important, and as such needs to be done properly. Therefore, researchers have identified some data interpretation methods to aid this process.

Methods involved in Data Interpretation

Data interpretation methods are how analysts help people make sense of numerical data that has been collected, analyzed and presented. Data, when collected in raw form, may be difficult for the layman to understand, which is why analysts need to break down the information gathered so that others can make sense of it. For example, when founders are pitching to potential investors, they must interpret data (e.g., market size, growth rate, etc.) for better understanding. There are 2 main methods in which this can be done, namely; quantitative methods and qualitative methods.

Qualitative Data Interpretation Method

The qualitative data interpretation method is used to analyze qualitative data, which is also known as categorical data. This method uses texts, rather than numbers or patterns to describe data. Qualitative data is usually gathered using a wide variety of person-to-person techniques, which may be difficult to analyze compared to the quantitative research method.

Unlike the quantitative data which can be analyzed directly after it has been collected and sorted, qualitative data needs to first be coded into numbers before it can be analyzed. This is because texts are usually cumbersome, and will take more time and result in a lot of errors if analyzed in its original state. Coding done by the analyst should also be documented so that it can be reused by others and also analyzed. There are 2 main types of qualitative data, namely; nominal and ordinal data. These 2 data types are both interpreted using the same method, but ordinal data interpretation is quite easier than that of nominal data.

In most cases, ordinal data is usually labelled with numbers during the process of data collection, and coding may not be required. This is different from nominal data that still needs to be coded for proper interpretation.

Quantitative Data Interpretation Method

The quantitative data interpretation method is used to analyze quantitative data, which is also known as numerical data. This data type contains numbers and is therefore analyzed with the use of numbers and not texts. Quantitative data are of 2 main types, namely; discrete and continuous data. Continuous data is further divided into interval data and ratio data, with all the data types being numeric.

Due to its natural existence as a number, analysts do not need to employ the coding technique on quantitative data before it is analyzed. The process of analyzing quantitative data involves statistical modelling techniques such as standard deviation, mean and median. Some of the statistical methods used in analyzing quantitative data are highlighted below:

1. Mean: The mean is a numerical average for a set of data and is calculated by dividing the sum of the values by the number of values in a dataset. It is used to get an estimate of a large population from the dataset obtained from a sample of the population.
2. Standard deviation: This technique is used to measure how well the responses align with or deviates from the mean. It describes the degree of consistency within the responses; together with the mean, it provides insight into data sets.
3. Frequency distribution: This technique is used to assess the demography of the respondents or the number of times a particular response appears in research. It is extremely keen on determining the degree of intersection between data points.

Some other interpretation processes of quantitative data not used as popularly as the previous three and uses quite hold a small niche includes:

- Regression analysis
- Cohort analysis
- Predictive and prescriptive analysis

Important points while collecting data for accurate data interpretation

- Identify the Required Data Type

  Researchers need to identify the type of data required for particular research. Is it nominal, ordinal, interval, or ratio data? The key to collecting the required data to conduct research is to properly understand the research question. If the researcher can understand the research question, then he can identify the kind of data that is required to carry out the research.

- Avoid Biases

  There are different kinds of biases a researcher might encounter when collecting data for analysis. Although biases sometimes come from the researcher, most of the biases encountered during the data collection process is caused by the respondent. There are 2 main biases, that can be caused by the President, namely; response bias and non-response bias. Researchers may not be able to eliminate these biases, but there are ways in which they can be avoided and reduced to a minimum.

  Response biases are biases that are caused by respondents intentionally giving wrong answers to responses, while non-response bias occurs when the respondents don't give answers to questions at all. Biases are capable of affecting the process of data interpretation.

- Use Close Ended Surveys

Although open-ended surveys are capable of giving detailed information about the questions and allow respondents to fully express themselves, it is not the best kind of survey for data interpretation. It requires a lot of coding before the data can be analyzed. Close-ended surveys, on the other hand, restrict the respondents' answer to some predefined options, while simultaneously eliminating irrelevant data. This way, researchers can easily analyze and interpret data.

Data Visualization

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made. Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data visualization is important for almost every career. It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders. It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Hence, visualization tools were a natural fit in these cases. Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning (ML) algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

Data visualization and big data

The increased popularity of big data and data analysis projects have made visualization more important than ever. Companies are increasingly using machine learning to gather massive amounts of data that can be difficult and slow to sort through, comprehend and explain. Visualization offers a means to speed this up and present information to business owners and stakeholders in ways they can understand. Big data visualization often goes beyond the typical techniques used in normal visualization, such as pie charts, histograms and corporate graphs. It instead uses more complex representations, such as heat maps and fever charts. Big data visualization requires powerful computer systems to collect raw data, process it and turn it into graphical representations that humans can use to quickly draw insights.

While big data visualization can be beneficial, it can pose several disadvantages to organizations. They are as follows:

- To get the most out of big data visualization tools, a visualization specialist must be hired. This specialist must be able to identify the best data sets and visualization styles to guarantee organizations are optimizing the use of their data.
- Big data visualization projects often require involvement from IT, as well as management, since the visualization of big data requires powerful computer hardware, efficient storage systems and even a move to the cloud.
- The insights provided by big data visualization will only be as accurate as the information being visualized. Therefore, it is essential to have people and processes in place to govern and control the quality of corporate data, metadata and data sources.

Techniques involved in visualization

In the early days of visualization, the most common visualization technique was using a Microsoft Excel™ spreadsheet to transform the information into a table, bar graph or pie chart. While these visualization methods are still commonly used, more intricate techniques are now available, including the following:

- infographics
- bubble clouds
- bullet graphs
- heat maps
- fever charts
- time series charts

Some other popular techniques are as follows.

- Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time.
- Area charts. This visualization method is a variation of a line chart; it displays multiple values in a time series -- or a sequence of data collected at consecutive, equally spaced points in time.
- Scatter plots. This technique displays the relationship between two variables. A scatter plot takes the form of an x- and y-axis with dots to represent data points.
- Treemaps. This method shows hierarchical data in a nested format. The size of the rectangles used for each category is proportional to its percentage of the whole. Treemaps are best used when multiple categories are present, and the goal is to compare different parts of a whole.
- Population pyramids. This technique uses a stacked bar graph to display the complex social narrative of a population. It is best used when trying to display the distribution of a population.

Importance of Visualization

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. The practice can also help businesses identify which factors affect customer behavior; pinpoint areas that need to be improved or need more attention; make data more memorable for stakeholders; understand when and where to place specific products; and predict sales volumes.

Other benefits of data visualization include the following:

- the ability to absorb information quickly, improve insights and make faster decisions;
- an increased understanding of the next steps that must be taken to improve the organization;
- an improved ability to maintain the audience's interest with information they can understand;
- an easy distribution of information that increases the opportunity to share insights with everyone involved;
- eliminate the need for data scientists since data is more accessible and understandable; and

- an increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

A little intro into the problem at hand

A semi hypothetical company 'FinMan' with various similar establishments popping from a simple name search in the google database, considering all these the most match to the problem initiated to us corroborates with either these three possible companies

1. Paarth FinMan, located at Kolkata, India.
2. Finman AG, located at Zollikon, ZÜRICH, Switzerland.

Since this name very closely matches along with the problem statement that was given in a job search hackathon by Analytics Vidya in the field of Data Science and also shared in the Kaggle community regarding the same field.

Given the task we are supposed to:

1. Study the data provided.
2. Find inconsistencies in the data.
3. Eventually clean the data by recommended methods.
4. Set up the required datatypes for each fields of data.
5. Fix the data to make it code readable.
6. Categorize or randomize data by requirement.
7. Regress logistically to get an interpretation of customer behavior.
8. Assess the metrics of the interpretation and check the efficacy of the method.
9. Make informed guess about potential buyers of health insurances from a fixed pool of buyers based on the knowledge gained of the customer interaction from the interpretation stated in the last point.
10. Provide a visual compilation of the buyer pool back to the customer('FinMan') with all leads and solutions.

Logistic Regression

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modelling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

Comparison to linear regression

Linear Regression and logistic regression can predict different things:

- Linear regression predictions are continuous (numbers in a range).
- Logistic regression predictions are discrete (only specific values or categories are allowed). We can also view probability scores underlying the model's classifications.

Types of logistic regression

- Binary (Pass/Fail)
- Multi (Cats, Dogs, Sheep)
- Ordinal (Low, Medium, High)

The Logit and Logistic Transformations

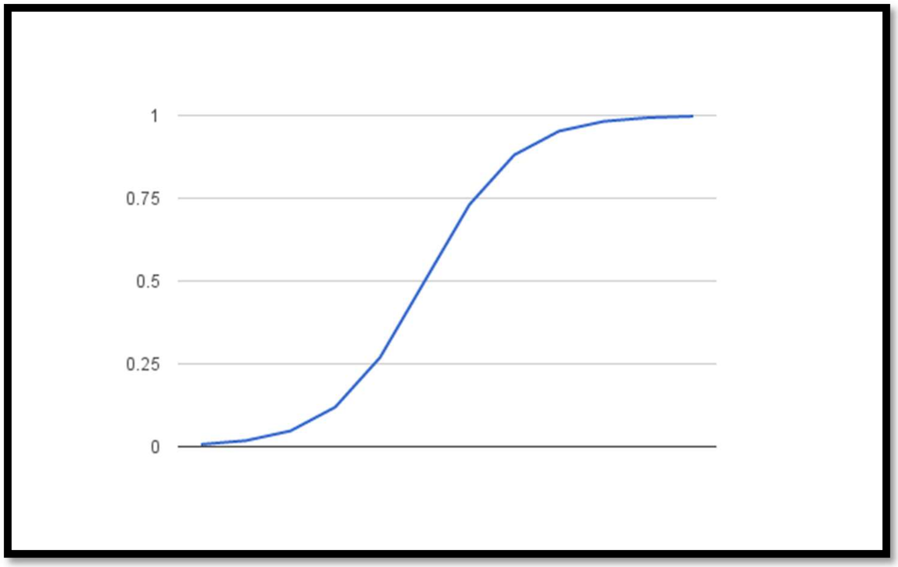
In multiple regression, a mathematical model of a set of explanatory variables is used to predict the mean of a continuous dependent variable. In logistic regression, a mathematical model of a set of explanatory variables is used to predict a logit transformation of the dependent variable. Suppose the numerical values of 0 and 1 are assigned to the two outcomes of a binary variable. Often, the 0 represents a negative response and the 1 represents a positive response. The mean of this variable will be the proportion of positive responses. If p is the proportion of observations with an outcome of 1, then 1-p is the probability of an outcome of 0. The ratio *p/(1-p)* is called the odds and the logit is the logarithm of the odds, or just log odds. Mathematically, the logit transformation is written

$$l = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

The logistic transformation is the inverse of the logit transformation. It is written

$$p = \text{logistic}(l) = \frac{e^l}{1+e^l}$$

This, gives us the equation for the sigmoid, which in perfect scenario gives us a curve of the likes described below.



The Log Odds Ratio Transformation

The difference between two log odds can be used to compare two proportions, such as that of males versus females. Mathematically, this difference is written

$$\begin{aligned} l_1 - l_2 &= \text{logit}(p_1) - \text{logit}(p_2) \\ &= \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) \\ &= \ln\left(\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_2}{1-p_2}\right)}\right) \\ &= \ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right) \\ &= \ln(OR_{1,2}) \end{aligned}$$

This difference is often referred to as the log odds ratio. The odds ratio is often used to compare proportions across groups. Note that the logistic transformation is closely related to the odds ratio. The reverse relationship is

$$OR_{1,2} = \mathrm{e}^{(l_1 - l_2)}$$

The Logistic Regression and Logit Models

In logistic regression, a categorical dependent variable Y having G (usually G = 2) unique values is regressed on a set of *p* independent variables $X_1, X_2, \ldots, X_p$. Since the names of these partitions are arbitrary, we often refer to them by consecutive numbers. That is, in the discussion below, Y will take on the values 1, 2, … *G*. Let,

$$\mathrm{X} = (X_1, X_2, \cdots, X_p)$$

$$B_g = \begin{pmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{pmatrix}$$

The logistic regression model is given by the *G* equations

$$\begin{aligned} \ln\left(\frac{p_g}{p_1}\right) &= \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \\ &= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g \end{aligned}$$

Here, p_g is the probability that an individual with values $X_1, X_2, \ldots, X_p$ is in outcome *g*. That is,

$$p_g = \Pr(Y = g \mid \mathrm{X})$$

Usually $X_1 \equiv 1$ (that is, an intercept is included), but this is not necessary. The quantities $P_1, P_2, \ldots, P_G$, represent the prior probabilities of outcome membership. If these prior probabilities are assumed equal, then the term $ln\ (P_g/P_1)$ becomes zero and drops out. If the priors are not assumed equal, they change the values of the intercepts in the logistic regression equation.

Outcome one is called the *reference value*. The regression coefficients β_{11}, β_{12}, …, β_{1p} for the reference value are set to zero. The choice of the reference value is arbitrary. Usually, it is the most frequent value or a control outcome to which the other outcomes are to be compared. This leaves G-1 logistic regression equations in the logistic model.

The β's are population regression coefficients that are to be estimated from the data. Their estimates are represented by b's. The β's represents unknown parameters to be estimated, while the b's are their estimates.

These equations are linear in the logits of p. However, in terms of the probabilities, they are nonlinear. The corresponding nonlinear equations are

$$\mathrm{p_g} = \mathrm{Prob}(Y = g \mid \mathrm{X}) = \frac{e^{\mathrm{XB}_g}}{1 + e^{\mathrm{XB}_2} + e^{\mathrm{XB}_3} + \cdots + e^{\mathrm{XB}_G}}$$

since $e^{XB_1} = 1$ because all of its regression coefficients are zero.

A note on the names of the models. Often, all of these models are referred to as logistic regression models. However, when the independent variables are coded as ANOVA type models, they are sometimes called *logit models*.

Another note about the interpretation of e^{XB} may be useful. Using the fact that $e^{a+b}=(e^a)(e^b)$, e^{XB} may be re-expressed as follows

$$\begin{aligned} e^{\mathrm{XB}} &= e^{\beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_p \mathrm{X}_p} \\ &= e^{\beta_1 \mathrm{X}_1} e^{\beta_2 \mathrm{X}_2} \cdots e^{\beta_p \mathrm{X}_p} \end{aligned}$$

This shows that the final value is the product of its individual terms.

Solving the Likelihood Equations

To improve notation, let

$$\begin{aligned} \pi_{gj} &= \mathrm{Prob}(Y = g \mid \mathrm{X}_j) \\ &= \frac{e^{\mathrm{X}_j \mathrm{B}_g}}{e^{\mathrm{X}_j \mathrm{B}_1} + e^{\mathrm{X}_j \mathrm{B}_2} + \cdots + e^{\mathrm{X}_j \mathrm{B}_G}} \\ &= \frac{e^{\mathrm{X}_j \mathrm{B}_g}}{\sum_{s=1}^{G} e^{\mathrm{X}_j \mathrm{B}_s}} \end{aligned}$$

The likelihood for a sample of N observations is then given by

$$l = \prod_{\mathrm{j}=1}^{\mathrm{N}} \prod_{g=1}^{G} \pi_{gj}^{\ y_{gj}}$$

where y_{gj} is one if the j^{th} observation is in outcome g and zero otherwise.

Using the fact that $\sum_{g=1}^{G} y_{gj} = 1$, the log likelihood, L, is given by

$$
\begin{aligned}
L = \ln(l) &= \sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(\pi_{gj}\right) \\
&= \sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(\frac{e^{X_jB_g}}{\sum_{s=1}^{G} e^{X_jB_s}}\right) \\
&= \sum_{j=1}^{N}\left[\sum_{g=1}^{G} y_{gj}X_jB_g - \ln\left(\sum_{g=1}^{G} e^{X_jB_g}\right)\right]
\end{aligned}
$$

Maximum likelihood estimates of the β's are those values that maximize this log likelihood equation. This is accomplished by calculating the partial derivatives and setting them to zero. The resulting likelihood equations are

$$
\frac{\partial L}{\partial \beta_{ik}} = \sum_{j=1}^{N} x_{kj}\left(y_{ig} - \pi_{ig}\right)
$$

for g = 1, 2, ..., G and k = 1, 2, ..., p. Actually, since all coefficients are zero for g= 1, the effective range of g is from 2 to G.

Because of the nonlinear nature of the parameters, there is no closed-form solution to these equations and they must be solved iteratively. The Newton-Raphson method as described in Albert and Harris (1987) is used to solve these equations. This method makes use of the information matrix, I(β), which is formed from the matrix of second partial derivatives. The elements of the information matrix are given by

$$
\frac{\partial^2 L}{\partial \beta_{ik}\partial \beta_{ik'}} = -\sum_{j=1}^{N} x_{kj}x_{k'j}\pi_{ig}\left(1 - \pi_{ig}\right)
$$

$$
\frac{\partial^2 L}{\partial \beta_{ik}\partial \beta_{i'k'}} = \sum_{j=1}^{N} x_{kj}x_{k'j}\pi_{ig}\pi_{i'g}
$$

The information matrix is used because the asymptotic covariance matrix of the maximum likelihood estimates is equal to the inverse of the information matrix. That is,

$$
V\left(\hat{\beta}\right) = I(\beta)^{-1}
$$

This covariance matrix is used in the calculation of confidence intervals for the regression coefficients, odds ratios, and predicted probabilities.

Interpretation of Regression Coefficients

The interpretation of the estimated regression coefficients is not as easy as in multiple regression. In logistic regression, not only is the relationship between X and Y nonlinear, but also, if the dependent variable has more than two unique values, there are several regression equations.

Consider the usual case of a binary dependent variable, Y, and a single independent variable, X. Assume that Y is coded so it takes on the values 0 and 1. In this case, the logistic regression equation is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Now consider impact of a unit increase in X. The logistic regression equation becomes

$$\begin{aligned}\ln\left(\frac{p'}{1-p'}\right) &= \beta_0 + \beta_1(X+1) \\ &= \beta_0 + \beta_1 X + \beta_1\end{aligned}$$

We can isolate the slope by taking the difference between these two equations. We have

$$\begin{aligned}\beta_1 &= \beta_0 + \beta_1(X+1) - (\beta_0 + \beta_1 X) \\ &= \ln\left(\frac{p'}{1-p'}\right) - \ln\left(\frac{p}{1-p}\right) \\ &= \ln\left(\frac{\frac{p'}{1-p'}}{\frac{p}{1-p}}\right) \\ &= \ln\left(\frac{odds'}{odds}\right)\end{aligned}$$

That is, β1 is the log of the ratio of the odds at *X+1* and *X*. Removing the logarithm by exponentiating both sides gives

$$e^{\beta_1} = \frac{odds'}{odds}$$

The regression coefficient β_1 is interpreted as the log of the odds ratio comparing the odds after a one unit increase in *X* to the original odds. Note that, unlike multiple regression, the interpretation of β_1 depends on the particular value of *X* since the probability values, the *p*'s, will vary for different *X*.

Binary X

When X can take on only two values, say 0 and 1, the above interpretation becomes even simpler. Since there are only two possible values of *X*, there is a unique interpretation for β_1 given by the log of the odds ratio. In mathematical terms, the meaning of β_1 is then

$$\beta_1 = \ln\left(\frac{odds(X=1)}{odds(X=0)}\right)$$

Multiple Independent Variables

When there are multiple independent variables, the interpretation of each regression coefficient becomes more difficult, especially if interaction terms are included in the model. In general, however, the regression coefficient is interpreted the same as above, except that the caveat 'holding all other independent variables constant' must be added. The question becomes, can the value of this independent variable be increased by one without changing any of the other variables. If it can, then the interpretation is as before. If not, then some type of conditional statement must be added that accounts for the values of the other variables.

Multinomial Dependent Variable

When the dependent variable has more than two values, there will be more than one regression equation. In fact, the number of regression equations is equal to one less than the number of outcomes. This makes interpretation more difficult because there are several regression coefficients associated with each independent variable. In this case, care must be taken to understand what each regression equation is predicting.

Problem Statement

The problem statement as was provided:

Project 1. Health Insurance Lead Prediction using Python

Your Client FinMan is a financial services company that provides various financial services like loan, investment funds, insurance etc. to its customers. FinMan wishes to cross-sell health insurance to the existing customers who may or may not hold insurance policies with the company. The company recommend health insurance to its customers based on their profile once these customers land on the website. Customers might browse the recommended health insurance policy and consequently fill up a form to apply. When these customers fill-up the form, their Response towards the policy is considered positive and they are classified as a lead.

Once these leads are acquired, the sales advisors approach them to convert and thus the company can sell proposed health insurance to these leads in a more efficient manner.

Now the company needs your help in building a model to predict whether the person will be interested in their proposed Health plan/policy given the information about:

Demographics (city, age, region etc.) Information regarding holding policies of the customer Recommended Policy Information

General Idea on inconsistencies from the Data Received

As per project job instructions we have three files at hand, two sets of data and a defining information depicting the data.

1. Train data
   - Filename: train_Df64byy.csv;
   - Records: 50882; Fields: 14;
   - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_1/Data/train_Df64byy.csv
2. Test data
   - Filename: test_YCcRUnU.csv
   - Records: 21805; Fields: 13;
   - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_1/Data/test_YCcRUnU.csv
3. Outline information on data fields
   - Filename: Metadata.docx
   - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_1/Data/Metadata.docx

According to observations with the help of spreadsheet software filtering tools for the datasets received we have the following problems.

Train data:

1. “Health indicator” field has 11691 blank or missing data records.
2. “Holding policy duration” and “Holding policy type” fields have 20251 missing or blank records.

3. "Holding policy duration" field has 4335 records with a value '14+' signifying a fact of 'more than 14' by units of time but the '+' character will create an error in runtime due to parsing issues, since it is not a numerical character.

Test data:

1. "Health indicator" field has 5027 blank or missing records.
2. "Holding policy duration" and "Holding policy type" fields have 8603 missing or blank records.
3. "Holding policy duration" field has 1892 records with a value '14+' signifying a fact of 'more than 14' by units of time but the '+' character will create an error in runtime due to parsing issues, since it is not a numerical character.

Additionally (Elephant of the room): The Train data has a field (viz., "Response") available for consideration which is completely missing from the test data.

Workaround to the Data problem

The most significant of the problem faced is the missing field of data producing the inconsistency from the test to train sets. As per suggestions from the guiding faculty the Train data as provided is to be split into a train – test pair with ratio as suggested by conventionally tested methods, and the Test data provided is to be produced up to the client to their requirements and fulfilling the question they had about predicting suggestions regarding the analysis of the sentiment of the potential buyer for a certain policy pitched towards them.

The missing record issue has a more standard solution suggested. For every missing value encountered, the nature of the missing record is to be assessed according to the field in question where it is found and then the following solution is to followed.

1. If the field of data is of categorical nature, then the missing records is to field with the modal value of the non-empty data records.
2. If the field of data is of numerical continuous nature then we need to check from the records, if from the non-empty records holds any outlier points.
    a. If outliers are present then the empty records are filled up with the median value of the non-empty records.
    b. In case of no outliers, the empty records are filled up with the mean value of the non-empty records.

Alternatively, the missing values may be handled via imputation algorithms like the KNN Imputer available from Scikit learning packages in python.

The issue with "Holding policy duration" field is to be handled by randomizing the records with the abnormality. Since we are given a cap of 20 for this field then for all cases of '14+' they are inferred to hold values between 15 to 20, hence, the records are to be randomized between these values.

Description of the data fields

The data fields described as a part of the problem at hand.

Field name	Description
ID	Unique key for each row
City_Code	Code of the city where customer lives
Region_Code	Equivalent to pin code
Accomodation_Type	Type of place to live
Reco_Insurance_Type	Recommended Insurance type
Upper_Age & Lower age	Eligible range of age for this insurance
Is spouse	Whether joint account holder is spouse or not
Health Indicator	Condition of health
Holding_Policy_Duration	If there is an existing policy, then what is the duration
Holding_Policy_Type	Existing policy type
Reco_Policy_Cat	Recommended policy category
Reco_Policy_Premium	Recommended policy premium amount
Response	Class 1 means customer will show interest in buying new policy.

Interpretation

<Interpretation>

Coding

Coding for this total project is divided into three sections for handling the complete job.

Type	Job nature
Operative tasks handlers	does each task separately from cleaning to decision creation
Support functions	mainly for small assessment of data or structure creators for container windows or UIs, works in

	tandem with Operative task handlers and engines depending on requirement
Engines	final engine or checkpoint from where the code starts running.

jkj

Conclusion and Inferences

From the Metrics obtained by calculations we can infer the following solutions

1.<EMPTY>=? TO FILL

Bibliography

The following materials were referenced for bringing the project to fruition.

- https://www.lotame.com/what-are-the-methods-of-data-collection/
- https://www.guru99.com/what-is-data-analysis.html
- https://www.investopedia.com/terms/d/data-analytics.asp
- https://web.archive.org/web/20201205235717/https://www.ibm.com/in-en/analytics/hadoop/big-data-analytics
- https://www.talend.com/resources/what-is-data-preparation/
- https://www.zarantech.com/blog/importance-of-data-science/
- https://www.trifacta.com/data-cleansing/
- https://www.sisense.com/glossary/data-cleaning/
- https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/
- https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825
- https://www.import.io/post/business-data-analysis-what-how-why/
- https://www.itl.nist.gov/div898/handbook/eda/eda.htm
- https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
- https://www.formpl.us/blog/data-interpretation
- https://searchbusinessanalytics.techtarget.com/definition/data-visualization
- https://www.tableau.com/learn/articles/data-visualization
- https://web.archive.org/web/20201101004343/https://machinelearningmastery.com/logistic-regression-for-machine-learning/
- http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html
- https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Logistic_Regression.pdf
- https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

- https://realpython.com/logistic-regression-python/
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

* * *

Complete collection of the project files is safely kept at

https://github.com/WolfDev8675/RepoSJX7/tree/Assign3_1

* * *