Exploratory Data Analysis and Visualization for Business Analytics

Chunhua Deming[1], Sreekanth Dekkati[2], Harshith Desamsetti[3]

[1]National University of Singapore, Singapore
[2]System Administrator, MUFG Bank, Arizona, USA
[3]School of Computer Science, Northern Illinois University, USA

ABSTRACT

The advantages of doing exploratory data analysis have been discussed in this article. We have gone through the typical data preprocessing procedures to understand the data and get it ready for modeling. In this research, we intend to become familiar with the most extensively utilized predictive modeling techniques and the fundamental principles underlying these techniques. Developing statistical or machine-learning models to generate predictions based on data is an example of predictive modeling, which is creating these models. We will use one typical example to make our research more tangible and demonstrate how the performance of various models varies when applied to the same data set. Exploratory data analysis will be covered in depth here, so prepare for that! In the following, we will discuss comprehending and validating the data.

Keywords: Data Analysis, Business Analytics, Data Visualization, EDA

INTRODUCTION

The process of translating data into insights for the purpose of improving company choices is known as business analytics. A number of different techniques, including data management and visualization, predictive modeling, data mining, forecasting simulation, and optimization, are utilized in the process of extracting insights from data. The term "exploratory data analysis," or EDA, refers to a technique that is utilized by data scientists to evaluate and investigate data sets as well as summarize the primary characteristics of such data sets. These techniques frequently involve the usage of data visualization approaches (Dekkati & Thaduri, 2017). It makes it simpler for data scientists to recognize patterns, recognize anomalies, test a hypothesis, or check assumptions by assisting them in determining the most effective way to alter data sources in order to obtain the answers they require.

In most cases, EDA is performed as a preliminary step before more formal statistical studies or modeling is attempted. There is no room for argument on this point; doing so will put you in a condition in which you will be overwhelmed by a great number of questions that demand solutions in order for you to come to a conclusion (Thodupunori & Gutlapalli, 2018). If you were an effective chieftain, the very first question you would ask is, "Who is in the cast and crew of the movie?" In addition to this, you would make it a habit to check out the movie's trailer on YouTube on a daily basis. In addition, you would learn the ratings and comments that were left by audience members after seeing the movie.

In the jargon of data scientists, whatever exploring steps you would do before finally purchasing popcorn for your family in the theater are referred to as "Exploratory Data Analysis." The term "Exploratory Data Analysis" refers to the crucial process of doing first investigations on data in order to uncover patterns, to spot anomalies, to test hypothesis, and to check assumptions with the use of summary statistics and graphical representations. This is a necessary step in every data analysis project (Thaduri et al., 2016).

Understanding the data first and making an effort to gain as much insight as possible from it is a recommended course of action. Before getting too mucky with the data, EDA is all about trying to make sense of what's already there. Even before we begin to construct any models, we will first become familiar with the core concepts of predictive analysis. Linear regression, a procedure utilized frequently in statistical modeling, served as the basis for our illustration of the concept. Even though linear regression is not a novel concept in traditional statistical analysis, as we will see, the emphasis that we place on it is rather different. As a predictive model, we make use of it. There is significantly less of an emphasis placed on traditional statistical notions like hypothesis. And a significantly increased focus on the accuracy of predictions. In the third unit, we will study how to create predictions for a result with two possible states. Something that cannot be handled by models of linear regression. In the fourth module, we delve a little deeper into predictive modeling and introduce two frequently used machine learning techniques, namely Trees and Neural Networks. Trees are a hierarchical representation of data, while neural networks are an artificial neural network. Because we will be using two distinct examples during the course of the study, we will be able to examine how different prediction models functioned on the same data set. It is our aim that by participating in this project, we will be able to build a strong foundation for predictive analytics. And develop an awareness of the irregular tension as well as the weaknesses. In addition to this, we shall see the connection between the various approaches. In spite of the fact that the prediction model has a great number of specifics and features, there are a few core concepts that are universal (Mandapuram, 2017a). With this newfound information, we will be better equipped to investigate predictive models beyond those that were given in this study. In this piece, we are going to talk about exploratory data analysis, which is one of the fundamental and necessary processes included in a data science project. A data scientist will spend roughly 70 percent of his time performing EDA on the dataset he is working with.

THE FOREMOST GOALS OF EDA

The term "Exploratory Data Analysis" (EDA) refers to the process of analyzing and investigating record sets with the goal of gaining an understanding of their major characteristics, discovering patterns, locating outliers, and determining the correlations between variables.

- Data Cleaning: EDA includes checking the data for errors, missing numbers, and inconsistencies in the data. It encompasses methods like as the imputation of data, the management of missing statistics, and the identification and elimination of outliers, among other things.
- Descriptive Statistics: EDA makes use of accurate data in order to identify the significant tendency, variability, and distribution of variables. In most cases, certain measurements such as the mean, median, mode, preferred deviation, range, and percentiles are utilized.

- Data Visualization: The EDA makes use of several visual techniques in order to graphically show the statistics. Histograms, box plots, scatter plots, line plots, heatmaps, and bar charts are all examples of visualizations that can be used to assist in recognizing patterns, trends, and correlations hidden within the data.
- Feature Engineering: EDA makes it possible to investigate different variables and the adjustments that may be made to them in order to generate new functionalities or draw relevant insights. Scaling, normalization, binning, encoding express variables, and the creation of interaction or derived variables are all examples of what can be included in feature engineering.
- Variable Dependencies and Relationships: EDA makes it possible to discover the variable dependencies and relationships between them. Techniques including as correlation analysis, scatter plots, and pass-tabulations are able to provide insights into the strength and direction of correlations between variables.
- Data Segmentation: EDA may include the process of splitting the information into substantial segments based solely on certain criteria or characteristics. This segmentation provides useful insights into distinct subgroups present within the data and may lead to more focused investigation.
- Generation of Hypotheses: EDA can be used to assist in the generation of hypotheses or research questions based just on the initial exploration of the data. It is helpful in forming the inspiration for further examination and model creation.
- Evaluation of the Information's Reliability and Quality EDA makes it possible to evaluate the accuracy and completeness of the data. In order to ensure that the information may be used for analysis, it is necessary to verify the records' integrity as well as their consistency and accuracy.

SIGNIFICANCE OF EXPLORATORY DATA ANALYSIS

This study covers exploratory data analysis before predictive model creation. Waste in, waste out. Bad data can confuse models. We explore common data difficulties and solutions. We sometimes use exploratory data analysis. Data summarization and realization. The data story will be revealed and acted upon. Real data is sometimes chaotic, with improper formatting, trailing spaces, duplication, empty rows, and abbreviated synonyms (Mandapuram, 2017b). Variations in scale, description, skewed distributions, outliers, and missing values. Each of these concerns might produce data analysis issues and needs consideration in exploratory data analysis. Some of these will be extensively explored in this module. Maintaining data quality, or integrity. Why so crucial? Data mistakes are common in clinical research, with rates between 2.3% and 26.9%. Sometimes a small mistake has serious implications. NASA lost a $125M Mars probe in 1999 because engineers forgot to convert English measurements to metric units. Simple Excel spreadsheet errors caused the $6 billion London whale trading loss. Data analysis generally begins with exploration. It involves data summarization and realization. It should precede modeling. Massive data can cause current problems in addition to data faults. A modern strategy may fail if a numerical variable has a long value. Common yield regression applies here. The garbage in, waste out effect can also affect huge data analysis. Data analysis on erroneous data might be deceptive (Gutlapalli, 2017a). Data exploration has many uses. First, it aids data comprehension. Sometimes called data understanding. Later data analysis can benefit from such intuition. They may even specify modeling methods or tools. Second, it helps sanity check data for reasonableness. Is formatted and sized correctly. The third benefit is identifying missing values outliers. Data investigation often concludes with a summary. The numeric summary and data exploration are included.

How should we explore data? Manual data review is easiest. For spreadsheet users, we may have manually reviewed data. Manual data review is too laborious. Excel spreadsheets can have 1 million rows. There are many approaches to analyze huge datasets. Examination of data samples comes first. Working with samples of data allows us to directly engage with raw data, even for big datasets, which can help us make sense of it. Finally, we must summarize the data using numerical summary or data validation.

DATA CLEANUP AND TRANSFORMATION

Data expiration, cleansing, and transformation were covered in this section. Here are some common data cleanup questions. Can character variables be valid? Numerical variables within range? Any missing values? Are values duplicated? Do ID variables have unique values? Are the dates correct? Do we need to merge data files? These questions must be checked and handled properly during data cleansing. Some of those challenges are easy to handle, but others are difficult. To handle missing values, one must understand the problem context and consider several options (Mandapuram, 2016). There are various commercial and open-source data cleansing tools. Popular open-source tools include OpenRefine and Data Wrangler. Many popular data exploration jobs benefit from the tools. We recommend visiting their websites for more. Data transformation usually involves applying a mathematical function to each data value. Centering and scaling a variable is maybe the most common data transformation. For statisticians, it's calculating each observed value's z-score. Each data value is decreased by the mean and divided by the standard deviation. Centering and scaling simplify and stabilize numerical methods immediately. Centralizing and scaling assures that dataset variables share a scale. Clustering, principal component analysis, and neural networks require or propose centering and scaling. The biggest negative is harder data interpretation. After centering and scaling, the data value measures each data point's uniqueness and standard deviation from the mean. Many more data transformations exist. Logarithm, square, square root, and inverse can express some of them. All of these transformations are polynomial save the first. Because they use data value polynomials (Gutlapalli, 2017b). Distinct problem contexts require distinct transformations. Choosing the proper one sometimes requires experimentation. Excel and other data analysis tools make these transformations easy.

The right data transformation can be automatically determined. Box-Cox transformation uses lambda. Lambda values transform differently. When lambda is 0, the transformation is logarithmic. It is a polynomial transformation with parameter lambda when not 0. A square transformation occurs when lambda = 2. Note that offsetting by 1 and dividing by lambda does not modify variable distribution.

Also, lambda = half is a square root transformation. The inverse transformation is lambda = -1. Lambda can be inferred from data and has numerous values. With this transformation, all previous transformations are covered. The transformations we've discussed work on one variable. Compare this to data reduction, which generates fewer variables. Data reduction uses fewer variables to capture most of the original variables' information (Desamsetti & Mandapuram, 2017). Principal component analysis, which finds weighted averages of variables to capture most data variance, is a popular data reduction method. New weighted averages are primary components. A few uncorrelated main components should capture most of the data's variance. We must scale the variables before doing principal component analysis. This prevents larger variables from dominating the major components.

DEALING WITH MISSING VALUES

Moving on to missing values in spirometry data processing is critical in this chapter. When rows or columns are empty, data sets often have missing values. Many missing values must be kept because we require enough data for useful analysis. Many leave them in and fill the values with smart estimations. In which scenario, minimize balances or distortions. Missing values may also reveal. Missing data points can be predictive (Gutlapalli, 2017c). Take a concrete example. The presentation exhibits newsstand newspaper sales data. Dates and daily sales are in the first and second columns, respectively. Curiously, March 27th sales are absent. This missing value might cause problems when modeling demands data values for each column or row. Example: linear regression.

Although removing all entries with missing values is easy, it can cause severe data loss. Datasets with several columns are significantly affected. Even with a tiny percentage of missing values per column, many rows will have at least one missing value. Too many rows will be eliminated. The non-plot on this slide shows sales data with a date x-axis and a copy count y-axis. Daily sales peak at 50 on March 28th. March 27th sales are unknown. That day has a missing value, hence the graph shows 0. The sales figures seem patterned. The value starts high, declines over the next few days, then rises again, repeating the process. This presentation shows the same graph with weekday data, which is useful. It appears from this graph that sales occur weekly. Sales are highest on Mondays, then progressively fall over the next several days but rise throughout the weekend. Note that the missing value on March 27th disrupts this pattern; without a right inference value, a real impression can be deceiving. What if we wanted to use this data for cost sales, a regular corporate operation, or to reduce cost by March 27th's zero sales value? This emphasizes the relevance of missing value preparation in predictive modeling. On March 27, what happened?

Many factors can produce missing values. Sometimes a value is missing because we forgot to record it. After some investigation, we realize that March 27th is Easter Sunday and the newsstand is closed. We know what caused the missing value on March 27th, but we need to decide what to do. Zero sales in our data collection will affect our sales projection. We cover several strategies to handle missing values. First, delete the data. As said, we may waste too much data, making it impractical. Second, assume or guess a value. We can substitute zero, average sales, or a smart approximation for the missing figure. Use last year's same-day sales to fill in the value. We can usually estimate the value using similar data points. Finally, we may categorize missing. Categorical data suits this method. Recall our example and assume a value for March 27th. Filling the value with zero, as done, is probably not a good idea. Sales would likely be higher if the newsstand weren't closed on Easter Sunday. Fill the value with average sales over all dates? That yields 37.23, which seems plausible. Since our data shows a weekly pattern, we can anticipate sales on the last Sunday.

ADDING AND REMOVING VARIABLES

This study discusses adding and removing variables. You can often locate more relevant data to add to the set. Internal data isn't always enough for sales forecasts. Collection of comparison and macroeconomic data is widespread. Sometimes a data set has duplicate or unhelpful columns. IT systems routinely store sales in different currencies.

However, only sales in one currency should be analyzed to avoid numeric issues. Additional variables might originate from many sources. First, data sets are generally gathered from public and online sources. Those sources typically provide extra information.

Dummy, converted, and arch-in-mean variables can be created from the data, and sometimes they are needed. We will briefly cover dummy variables, transformed variables, and arch-in-mean terms later in the study. Dummy variables are often created from a single variable with a few fixed values. Each value is a category and the variable is categorical. Dummy variables can be created from categorical variables with m categories, where m is an integer. Dummy variables normally have values of 0 or 1, with 1 indicating yes and 0 no.

For instance, the left table shows single-family homes, townhouses, and condos. Three categories and m=3. Two timing variables, D1 and D2, can indicate if it's a single-family home or a townhouse. D1 and D2 indicating 0 signify condos, which are neither single-family nor townhouses. When there are several categories for a categorical variable, this method can create many dummy variables, which can be troublesome when observations are sparse. Combining similar categories reduces the number of categories.

In the original data set, there are additional categories besides single-family home, townhouse, and condo. Some categories may be combined depending on our findings. Condos, coops, multi-family homes, and mobile homes are examples of others. Thus, we use two dummy variables instead of five if we don't combine categories. Category combinations aren't always easy and should be picked carefully. Combining sparse categories is typical, although the problem context may affect it. This function is built into many software products. As mentioned, we may want to eliminate variables from our data. To simplify model and analysis is the initial motivation. Some variables may not provide relevant information and should not be analyzed. Sometimes two variables have the identical information, thus we should maintain one. More parsimonious and easier-to-interpret models result from eliminating variables and including fewer variables. Some methods fail when a variable has a degenerate distribution, meaning it only takes one value (Gutlapalli, 2016b). Degenerate distribution variables must be deleted for improved performance. Exploratory data analysis calls degenerative variables zero variance variables. If all students in a high school class are born in the same year, the first year is a zero variance variable and should not be analyzed because it may produce numeric problems. Near-zero variance variables with a short data range are more common. Again, if all students in the class are born in the same year except one, the first year has near-zero variance. Keeping zero or near-zero variance variables in regression analysis can produce cardinality concerns, where coefficient estimates might alter unpredictably with tiny data changes, which is undesirable.

EXPLORATORY DATA ANALYSIS TOOLS

The following is a list of some of the most prevalent data science tools that are used to develop an EDA:

- Python is a computer programming language that is interpreted, object-oriented, and has dynamic semantics. Its high-level data structures that are built-in, along with dynamic typing and dynamic binding, make it a very attractive choice for the rapid building of applications, as well as for usage as a scripting or glue language to bring together existing components. Python and EDA can be used together to find missing values in a data collection, which is crucial so that we can decide how to treat missing values for machine learning. This can be done by combining the two tools.
- R is a free and open-source programming language that is used for statistical computation and graphics. R is maintained by the R Foundation for Statistical computation. The programming language R is utilized extensively in the field of data science by statisticians for the purpose of making statistical observations and doing data analysis.

GOOD DATA VISUALIZATION

Data exploration requires data manipulation, as we saw before. It can also aid data processing and interpretation. Analyses can end with data modification. Data manipulation-only analysis is powerful. It's constructed on raw data without statistical assumptions, unlike many statistical models (Desamsetti, 2016). Model-free analysis is important in big data analytics when we have a lot of data. I want to briefly explore graphical design concepts, a fascinating yet broad topic.

Stephen Few says our visual design goal is to showcase crucial content for readers (Gutlapalli, 2016a). Organises everything for clarity and conveys the story in the best order. Even when constructing simple graphs, one must follow graphical design principles. First, remove as much non-data-ink as feasible and use the rest to support. Thus, data validation should focus on data, not non-data. Organizing vital data-ink is important too.

Finally, when constructing data validations, we should modify, explore, and repeat.

Few thinks we should keep few to quantity. Both left and right plots display the same data. Both represent Amtrak's five-year passenger total. Ridership appears to be growing on the left but flat on the right.

We've mentioned pie charts and bar graphs for categorical data. Pie charts are popular, although few advises against them. His key point is that pie slices and quantities are hard to correlate.

It compels us to read and compare pie portions, which many find difficult. A bar graph makes bar data easier to compare.

When there are several categories, it gets worse. Indeed, inventive pie charts are among the worst web visuals.

Do not use points for time series data. We should highlight individual values with bars and trends using lines.

All three graphs display the same data, but they have distinct perspectives. Which one do we prefer?

Data validation software has grown rapidly in recent years, with new capabilities constantly released.

I list 3D validation software. Tableau dominates data validation software. Also popular are Click and Spotfire.

CONCLUSION

Exploratory data analysis is a straightforward categorization procedure that is typically carried out through the use of visual methods. It is a method for examining sets of data in order to summarize the most important aspects of those sets. When we are attempting to construct a model using machine learning, we need to be rather certain about whether or not the data we are using makes sense. Exploratory data analysis, also known as EDA, is the process of examining data by making use of basic tools from the field of statistics, such as basic charting tools. These are some of the more fundamental charts utilized in EDA. Read carefully and make sure you comprehend all that's going on in the plot at all times. It is never a smart idea to skip EDA for a project that involves machine learning.

REFERENCES

Dekkati, S., & Thaduri, U. R. (2017). Innovative Method for the Prediction of Software Defects Based on Class Imbalance Datasets. *Technology & Management Review*, *2*, 1–5. Retrieved from https://upright.pub/index.php/tmr/article/view/78

Desamsetti, H. (2016). Issues with the Cloud Computing Technology. *International Research Journal of Engineering and Technology (IRJET), 3*(5), 321-323.

Desamsetti, H., & Mandapuram, M. (2017). A Review of Meta-Model Designed for the Model-Based Testing Technique. *Engineering International*, *5*(2), 107–110. https://doi.org/10.18034/ei.v5i2.661

Gutlapalli, S. S. (2016a). An Examination of Nanotechnology's Role as an Integral Part of Electronics. *ABC Research Alert*, *4*(3), 21–27. https://doi.org/10.18034/ra.v4i3.651

Gutlapalli, S. S. (2016b). Commercial Applications of Blockchain and Distributed Ledger Technology. *Engineering International*, *4*(2), 89–94. https://doi.org/10.18034/ei.v4i2.653

Gutlapalli, S. S. (2017a). Analysis of Multimodal Data Using Deep Learning and Machine Learning. *Asian Journal of Humanity, Art and Literature*, *4*(2), 171–176. https://doi.org/10.18034/ajhal.v4i2.658

Gutlapalli, S. S. (2017b). The Role of Deep Learning in the Fourth Industrial Revolution: A Digital Transformation Approach. *Asian Accounting and Auditing Advancement*, *8*(1), 52–56. Retrieved from https://4ajournal.com/article/view/77

Gutlapalli, S. S. (2017c). An Early Cautionary Scan of the Security Risks of the Internet of Things. *Asian Journal of Applied Science and Engineering*, *6*, 163–168. Retrieved from https://ajase.net/article/view/14

Mandapuram, M. (2016). Applications of Blockchain and Distributed Ledger Technology (DLT) in Commercial Settings. *Asian Accounting and Auditing Advancement, 7*(1), 50–57. Retrieved from https://4ajournal.com/article/view/76

Mandapuram, M. (2017a). Application of Artificial Intelligence in Contemporary Business: An Analysis for Content Management System Optimization. *Asian Business Review*, *7*(3), 117–122. https://doi.org/10.18034/abr.v7i3.650

Mandapuram, M. (2017b). Security Risk Analysis of the Internet of Things: An Early Cautionary Scan. *ABC Research Alert*, *5*(3), 49–55. https://doi.org/10.18034/ra.v5i3.650

Thaduri, U. R., Ballamudi, V. K. R., Dekkati, S., & Mandapuram, M. (2016). Making the Cloud Adoption Decisions: Gaining Advantages from Taking an Integrated Approach. *International Journal of Reciprocal Symmetry and Theoretical Physics*, *3*, 11–16. https://upright.pub/index.php/ijrstp/article/view/77

Thodupunori, S. R., & Gutlapalli, S. S. (2018). Overview of LeOra Software: A Statistical Tool for Decision Makers. *Technology & Management Review, 3*(1), 7–11.

--0--