**CHAPTER 9**

**DATASET PREPARATION**

When we talk about data, we usually think of some large datasets with huge numbers 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. In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm. A dataset can be viewed as a collection of data objects, which are often also called records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called variables, characteristics, fields, attributes, or dimensions.

**7.1 Loading Libraries**

We'll make use of the following packages:

1. numpy and pandas is what we'll use to manipulate our data
2. matplotlib.pyplot and seaborn will be used to produce plots for visualization 85
3. util will provide the locally defined utility functions that have been provided for this assignment.
4. **Exploratory Data Analysis**

Exploratory data analysis (EDA) is performed in order to gain a preliminary understanding and allow us to get acquainted with the dataset. In a typical data science project, one of the first things that I would do is “eyeballing the data” by performing EDA so as to gain a better understanding of the data.

Three major EDA approaches that I normally use includes:

1. **Descriptive statistics** **—** Mean, median, mode, standard deviation
2. **Data visualisations** **—** Heat maps (discerning feature intra-correlation), box plot (visualize group differences), scatter plots (visualize correlations between features), principal component analysis (visualize distribution of clusters presented in the dataset), etc.
3. **Data shaping/ cleaning —** Pivoting data, grouping data, filtering data, etc.
   1. **Descriptive Statistics**

In Descriptive Statistics you are describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables. There are four major types of descriptive statistics:

1. Measures of Frequency: \* Count, Percent, Frequency.
2. Measures of Central Tendency: \* Mean, Median, and Mode.
3. Measures of Dispersion or Variation: \* Range, Variance, Standard Deviation.
4. Measures of Position: \* Percentile Ranks, Quartile Ranks.

**7.3 Dataset Visualization**

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. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions. Our eyes are drawn to colours and patterns. We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be. As the “age 87 of Big Data” kicks into high-gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. Data visualization helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers. A good visualization tells a story, removing the noise from data and highlighting the useful information.

However, it’s not simply as easy as just dressing up a graph to make it look better or slapping on the “info” part of an infographic. Effective data visualization is a delicate balancing act between form and function. The plainest graph could be too boring to catch any notice or it makes a powerful point; the most stunning visualization could utterly fail at conveying the right message or it could speak volumes. The data and the visuals need to work together, and there’s an art to combining great analysis with great storytelling. It’s hard to think of a professional industry that doesn’t benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

While we’ll always wax poetically about data visualization (you’re on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it’s also one of the most useful professional skills to develop. The better you can convey your points visually, whether in a dashboard or a slide deck, the better you can leverage that information. The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs the who, what, when, where, and how. While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling. Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen. With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present. But data visualization is not only important for data scientists and data analysts, it is necessary to understand data visualization in any career. Whether you work in finance, 88 marketing, tech, design, or anything else, you need to visualize data.

* 1. **Dataset Cleaning**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.

* + 1. **How do you clean data?**

While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to map out a framework for your organization.

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. Deduplication is one of the largest areas to be considered in this process.

Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze. For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performant dataset.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabelled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analysed as the same category.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analysing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on.

Remember: just because an outlier exists, doesn’t mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data

You can’t ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

1. As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
2. As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
3. As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Inconsistent values

We know that data can contain inconsistent values. Most probably we have already faced this issue at some point. For instance, the ‘Address’ field contains the ‘Phone number’. It may be due to human error or maybe the information was misread while being scanned from a handwritten form. It is therefore always advised to perform data assessment like knowing what the data type of the features should be and whether it is the same for all the data objects. We had age as a string type in the form “66Y” where “Y” needs to be eliminated and converted to integer type for processing.

Step 6: Duplicate values

A dataset may include data objects which are duplicates of one another. It may happen when the same person submits a form more than once. The term deduplication is often used to refer to the process of dealing with duplicates. In most cases, the duplicates are removed so as to not give that particular data object an advantage or bias, when running machine learning algorithms.

Step 7: Validate and QA

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

1. Does the data make sense?
2. Does the data follow the appropriate rules for its field?
3. Does it prove or disprove your working theory, or bring any insight to light?
4. Can you find trends in the data to help you form your next theory?
5. If not, is that because of a data quality issue?

False conclusions because of incorrect or “dirty” data can inform poor business strategy and decision-making. False conclusions can lead to an embarrassing moment in a reporting meeting when you realize your data doesn’t stand up to scrutiny.

Before you get there, it is important to create a culture of quality data in your organization. To do this, you should document the tools you might use to create this culture and what data quality means to you.

**7.4.2 Components of quality data**

Determining the quality of data requires an examination of its characteristics, then weighing those characteristics according to what is most important to your organization and the application(s) for which they will be used.

1. **Characteristics of quality data**
2. **Validity**: The degree to which your data conforms to defined business rules or constraints.
3. **Accuracy:** Ensure your data is close to the true values.
4. **Completeness:** The degree to which all required data is known.
5. **Consistency:** Ensure your data is consistent within the same dataset and/or across multiple data sets.
6. **Uniformity:** The degree to which the data is specified using the same unit of measure.

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

1. Removal of errors when multiple sources of data are at play.
2. Fewer errors make for happier clients and less-frustrated employees.
3. Ability to map the different functions and what your data is intended to do.
4. Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
5. Using tools for data cleaning will make for more efficient business practices and quicker decision-making.
   1. **Feature Sampling**

Sampling is a very common method for selecting a subset of the dataset that we are analyzing. In most cases, working with the complete dataset can turn out to be too expensive considering the memory and time constraints. Using a sampling algorithm can help us reduce the size of the dataset to a point where we can use a better, but more expensive, machine learning algorithm. The key principle here is that the sampling should be done in such a manner that the sample generated should have approximately the same properties as the original dataset, meaning that the sample is representative. This involves choosing the correct sample size and sampling strategy. Simple Random Sampling dictates that there is an equal probability of selecting any particular entity. It has two main variations as well:

1. Sampling without Replacement: As each item is selected, it is removed from the set of all the objects that form the total dataset.
2. Sampling with Replacement: Items are not removed from the total dataset after 93 getting selected. This means they can get selected more than once. Although simple random sampling provides two great sampling techniques, it can fail to output a representative sample when the dataset includes object types which vary drastically in ratio. This can cause problems when the sample needs to have a proper representation of all object types, for example, when we have an imbalanced dataset. An imbalanced dataset is one where the number of instances of a classes are significantly higher than another classes, thus leading to an imbalance and creating rarer classes.
   1. **Dimensionality Reduction**

Most real world datasets have a large number of features. For example, consider an image processing problem, we might have to deal with thousands of features, also called dimensions. As the name suggests, dimensionality reduction aims to reduce the number of features - but not simply by selecting a sample of features from the feature-set, which is something else — Feature Subset Selection or simply Feature Selection. Conceptually, dimension refers to the number of geometric planes the dataset lies in, which could be so high that it cannot be visualized with pen and paper. More the number of such planes, more is the complexity of the dataset.

**7.6.1 The Curse of Dimensionality**

This refers to the phenomena that generally data analysis tasks become significantly harder as the dimensionality of the data increases. As the dimensionality increases, the number planes occupied by the data increases thus adding more and more sparsity to the data which is difficult to model and visualize. A representation of how principal components can be visualized. Representation of components in different spaces. What dimension reduction essentially does is that it maps the dataset to a lower-dimensional space, which may very well be to a number of planes which can now be visualized, say 2D. The basic objective of techniques which are used for this purpose is to reduce the dimensionality of a dataset by creating new features which are a combination of the old features. In other words, the higher-dimensional feature-space is mapped to a lower-dimensional feature-space. Principal Component Analysis and Singular Value Decomposition are two widely accepted techniques. Data Analysis algorithms work better if the dimensionality of the dataset is lower. This is mainly because irrelevant features and noise have now been eliminated. The models which are built on top of lower dimensional data are more understandable and explainable.

**7.6.2 Feature Encoding**

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding. As mentioned before, the whole purpose of data preprocessing is to encode the data in order to bring it to such a state that the machine now understands it. Feature encoding is basically performing transformations on the data such that it can be easily accepted as input for machine learning algorithms while still retaining its original meaning. Data frame analytics automatically performs feature encoding. The input data is pre-processed with the following encoding techniques:

1. one-hot encoding: Assigns vectors to each category. The vector represents whether the corresponding feature is present (1) or not (0).
2. target-mean encoding: Replaces categorical values with the mean value of the target variable.
3. frequency encoding: Takes into account how many times a given categorical value is present in relation with a feature.

When the model makes predictions on new data, the data needs to be processed in the same way it was trained. Machine learning model inference in the Elastic Stack does this automatically, so the automatically applied encodings are used in each call for inference. Feature importance is calculated for the original categorical fields, not the automatically encoded features.