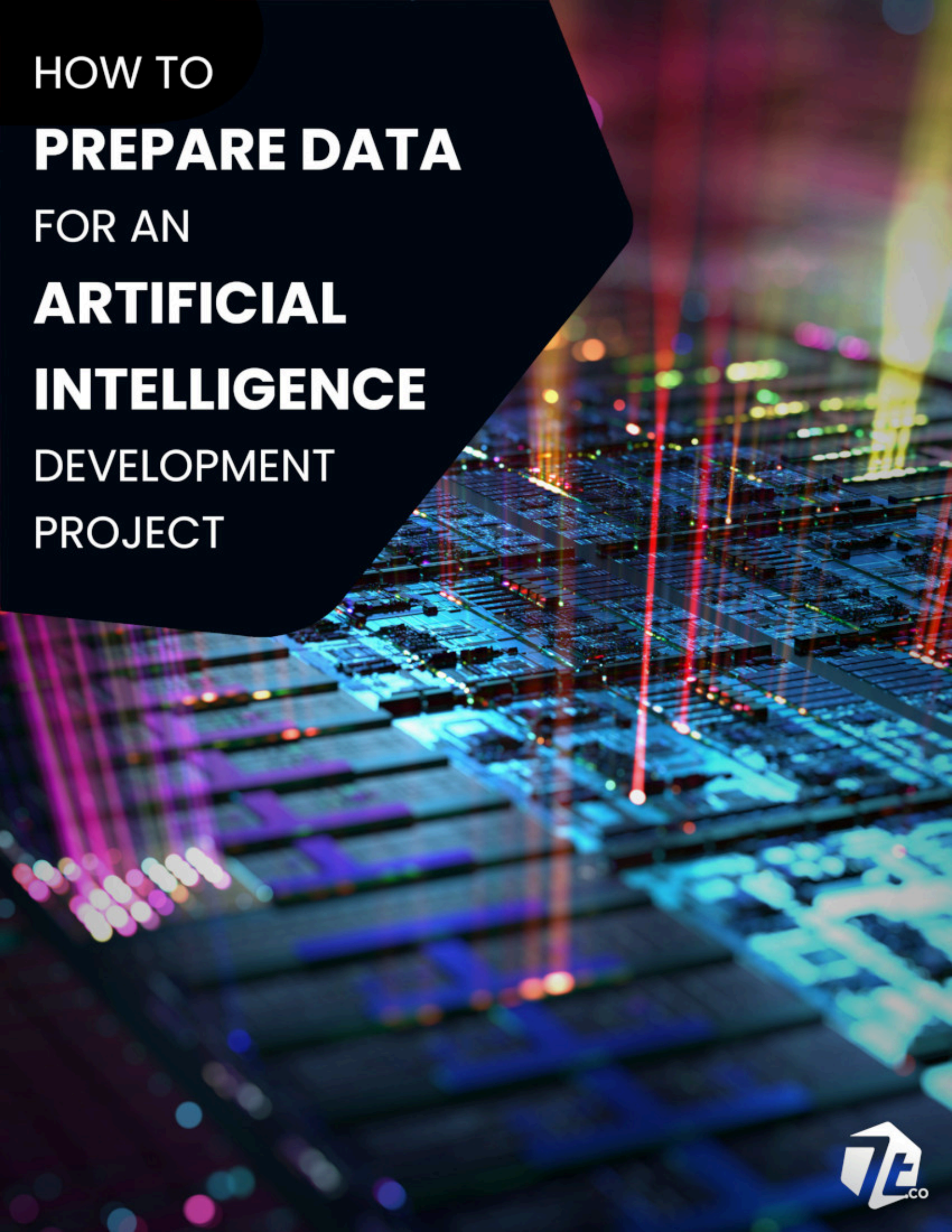
HOW TO **PREPARE DATA** FOR AN **ARTIFICIAL INTELLIGENCE** DEVELOPMENT PROJECT



How to Prepare Data for an
Artificial Intelligence Development Project

INDEX

Introduction 2
Project Objectives and Data Requirements 2
Defining Challenges and Goals ..... 2 - 3
Aligning Data With Your Organization's Objectives 3
How is Data Classified? Structured, Unstructured and Semi-Structured ..... 3 - 4
Stakeholder Collaboration for Successful AI Development 4
Data Collection and Integration ..... 4 - 5
Data Labeling During the Data Preparation Process ..... 5 - 6
The Data Cleaning Process ..... 6 - 7
The Data Transformation Process ..... 7 - 8
Exploratory Data Analysis (EDA) ..... 8 - 9
Feature Selection 9
Data Splitting ..... 9 - 11
Data Augmentation 11
Data Pipelines 12
Data Prep Drives the Most Innovative, Lucrative AI Projects ..... 12 -13
Contact 7T 14

Machine learning-driven AI has emerged as a game-changer in the world of technology, with virtually every company wondering, "How can I use AI for my business?" And while new uses for AI emerge on a daily basis, many of these use cases fall into the more novel realm of using AI for the sake of using AI.

Truly transformative AI requires innovation. And innovation is fueled by data. This begs the question: Is your data ready to propel your artificial intelligence development project forward and beyond expectations?

Project Objectives And Data Requirements

To position your organization for success with a machine learning-driven AI development project, you'll need to begin by addressing the following points.

- **Define Project Goals** - Clarify what business problem the AI/ML model is solving and identify measurable objectives that will allow you to determine if you've achieved success.

- **Align Data With Goals** - Identify specific data types (e.g. structured or unstructured data) that are needed to meet the stated objectives.

- **Stakeholder Collaboration** -Involve your company's teams and leaders to ensure the data requirements match real-world expectations and constraints.

Defining Challenges And Goals

Before a single byte of data is considered, an organization's stakeholders must convene to determine what pain point, challenge or problem exists within their organization. Take some time to clearly identify and articulate the issue(s).

Once you have a clearly-articulated problem, it's time to find the ideal solution. While many immediately go to a solution that involves machine learning-powered artificial intelligence, it's important that you perform a thorough exploration of the available options. AI is trendy — and for good reason — but it's not the perfect solution for every problem. If you're uncertain of what other high-tech alternatives may exist, consider consulting with a company that deals in more than just ML and AI. Seek a company — such as 7T — that specializes in Digital Transformation; this way, you can learn about other technologies that may be well-suited for addressing your

challenges, pain points or objectives. In many cases, the ideal solution involves multiple technologies, such as AI, ML, process automation and cloud computing.

Once an organization has a clear idea of the problem and the ideal high-tech solution, it's time to create specific, measurable objectives, each associated with a timeframe. This way, you can monitor progress to ensure that your AI development project remains on the path toward success.

Aligning Data With Your Organization's Objectives

Stated simply, good data preparation is essential for success in today's techsphere, whether it's ML-driven artificial intelligence or another innovative, cutting-edge technology. Unfortunately, many business leaders quickly come to realize that their data is scattered across disparate systems and in forms that limit its usefulness within the context of an AI development project.

The solution? You must perform a complete and systematic review of your data, its structure and other attributes. Then, align your data with your objectives. Determine what data types are required to meet those objectives. This requires data classification.

How is Data Classified?

Structured, Unstructured & Semi-Structured Data

Data can be classified in one of three ways:

- **Unstructured Data –** Unstructured data cannot be neatly organized into distinct, precise categories on a spreadsheet or similar.. It comes in many forms, such as images, emails, social posts, PDFs, audio and video files and so forth. Unstructured data is more erratic, making it more difficult to analyze, query and otherwise leverage that data in a meaningful way.

- **Structured Data** - Structured data can be neatly organized into distinct fields, with information such as dates and times, names, monetary figures, product identification numbers and transaction numbers, amongst others. Structured data can be easily queried and analyzed.

• **Semi-Structured Data** - Semi-structured data has traits of both structured and unstructured data. It may contain organizational elements like tags or metadata that provide structure but lack the rigid framework of traditional structured data. Common examples include JSON, XML, and NoSQL databases. While some parts of the data can be visualized or organized in a spreadsheet, other parts (like multimedia content) do not fit neatly into rows and columns.

Structured data is often quantitative, while unstructured data is typically — but not always — qualitative. Structured data is commonly stored in data warehouses, whereas unstructured data is typically stored in a data lake. Structured data exists in predefined formats, whereas unstructured data comes in a variety of different formats.

Stakeholder Collaboration For Successful AI Development

Stakeholder collaboration is essential for success, so take the time to reach out to team leaders, department heads and others in a leadership role within your organization. Ideally, these stakeholders ought to be involved in the early stages of your AI development process,, as you pinpoint challenges, pain points and future objectives.

An organization's stakeholders are well-positioned to evaluate your data requirements and to ensure that they align with real-world expectations and constraints. These insights are essential as an organization's leadership works to establish data preparation plans and to refine the project plan so that it remains on a path toward success.

Data Collection And Integration

A solid data collection and integration plan is critical to the success of a machine learning-driven AI project. There are four basic steps that fall into this stage of the project.

• **Identifying Data Sources** - What data does your organization generate? Pinpoint any and all data sources, including:

- Databases;
- APIs;
- Data Logs; and
- Third-Party Datasets,

Data is commonly spread amongst multiple sources, especially in the case of third-party data, so it's important to involve leaders from all departments and divisions. These individuals may have insights into important but little-known data sources that ought to be included in the project.

• **Data Integration –** Data integration is an important part of the preparation process. Explore the different options for merging and aligning the data that originates from disparate sources. From ETL processes to database joins, the best option varies according to the data classification, amongst other factors.

• **Handling Heterogeneous Data –** In this stage of the data preparation process, you'll need to address issues related to creating homogeneous data. With heterogeneous data, you may see differences in data types, data formats, data structure and scales of data. This will include structured vs unstructured data and batch processing data vs. streaming data.

• **Data Privacy –** Data privacy must be evaluated, especially if your organization is subject to stringent privacy regulations, such as those that are seen in the healthcare field. Healthcare clinics, hospitals and others in this industry must comply with strict, far-reaching patient data privacy requirements under HIPAA. Similarly, any company that does business with EU citizens must be GDPR-compliant.. You must ensure that your data collection, data transmission and data storage practices are in line with relevant privacy laws. An organization must also apply anonymization and / or encryption measures to protect data during the integration process.

Data Labeling During The Data Preparation Process

IBM describes data labeling as a mechanism that "provides users, teams and companies with greater context, quality and usability. More specifically, you can expect more precise predictions. Accurate data labeling ensures better quality

assurance within machine learning algorithms, allowing the model to train and yield the expected output."

What's more, labeling makes data more "usable," making it possible to " improve [the] usability of data variables within a model. For example, you might reclassify a categorical variable as a binary variable to make it more consumable for a model. Aggregating data in this way can optimize the model by reducing the number of model variables or enable the inclusion of control variables.

In the realm of machine learning — a technology that is critical to multiple forms of AI — data labels are "critical in supervised machine learning as targets for optimizing predictive models using historical input and output data, making them the ground truth to train models for classification and regression problems."

- **Supervised vs Unsupervised Learning -** The key difference between supervised and unsupervised learning is the data type. Supervised learning requires labeled data, whereas unsupervised learning uses unlabeled data. WIth this in mind, you'll need to determine what labeling is necessary to achieve success with supervised learning tasks.

- **Manual Labeling vs Automated Labeling** - There are two basic options for data labeling. Manual labeling allows for a high level of accuracy and overall quality. Meanwhile, automated labeling is favored for its speed and cost efficiency, although there is a loss in overall quality / accuracy.

- **Quality Control Considerations -** Accurate, consistent labeling is an essential quality control consideration, as it helps you to avoid issues such as model bias.

Notably, in some cases, particularly with raw data, the data labeling process may be more effective after cleaning and transformation. For example, if raw imaging data contains noise or artifacts, cleaning it before labeling can ensure that the labeling process focuses on relevant features, resulting in better-quality labels.

The Data Cleaning Process

Data cleaning is a very important part of the data preparation process, as all subsequent tasks are dependent upon "clean" datasets. Data cleaning consists of several tasks, including the following.

• **Handle Missing Values** - This segment of the data cleaning process involves imputation, deletion and other advanced techniques for managing missing data.

• **Remove Duplicates and Inconsistencies** - In this step, you ensure data integrity by standardizing formats and eliminating duplicate data.

• **Outlier Management** - Outlier management involves approaches for detecting and managing data outliers, including removal, transformation and separate analysis.

• **Standardizing Data Formats** - By standardizing data formats, you ensure consistency across the entire dataset by converting data so it has a common format and a common unit of measurement.

• **Data Accuracy Validation** - This task involves verifying the quality, authenticity and overall accuracy of a dataset. As such, data validation is widely considered a form of data cleaning.

• **Data Profiling** - Data profiling involves sorting, cleansing and analysis to ensure that the dataset is valid and ready for a more comprehensive analysis.

• **Continual Monitoring and a Feedback Loop** - By creating a feedback loop, you can verify that a particular dataset has been cleaned. Continual monitoring allows for ongoing refinements in your datasets.

Monitoring for errors and establishing an error log are common parts of the data cleaning process. By keeping a record of what errors have occurred and where those errors originated, you can identify and fix incorrect or corrupt data.

It's also prudent to establish a data cleaning workflow. Once established, this process can be shared with fellow team members who can then take part in this aspect of the data cleaning process.

The Data Transformation Process

The data transformation process involves converting raw data into a structured, uniform format that can be queried, analyzed and leveraged for data-driven decision-making. Data transformation involves several steps, including the following.

• **Data Normalization and Scaling** - Normalization reduces data redundancy and improves overall integrity through data reorganization and standardization. This process leads to data that's easy to query and analyze. By adjusting the scale of numerical features, you can prevent features with larger ranges from disproportionately dominating other features during the model training process. By scaling the features, each one contributes more equally to the final model. As a result, optimization algorithms converge more quickly during model training by starting all feature values in a consistent range. Some of the most common normalizing and scaling techniques include min-max scaling and Z-score normalization.

• **Encoding Categorical Variables** - One-hot coding is one of the most popular methods for encoding categorical variables. One-hot encoding involves the creation of dummy variables. This technique is used with categorical variables whereby order doesn't matter. You can use one-hot encoding with nominal features. With this technique, one new variable is created for every categorical feature.

• **Data Type Conversion** - Data type conversion involves the process of changing a value from one data type to another. Data type conversion is used when working with different data sources or when transforming data to prepare for analysis. This method can be used to convert data types — such as string to integer — to ensure compatibility with a machine learning algorithm.

• **Handling Imbalanced Datasets** - There are a number of techniques for handling imbalanced datasets. You can apply techniques such as oversampling, undersampling or the synthetic minority oversampling technique (SMOTE) to address class imbalance. There's also downsampling and upweighting the majority class. In this context, downsampling involves training on a disproportionately low subset of majority class examples. On the other hand, upweighting refers to the practice of adding an example weight to the downsampled class that's equal to the factor that was used to downsample the dataset.

• **Time Series-Specific Transformations** - Time series-specific transformations involve the creation of a log of features to incorporate historical data and calculate rolling statistics in an effort to capture trends and seasonal patterns in time series data.

• **Other Forms of Data Transformation** - Additional forms of data transformation include data source identification, data gathering, mapping modifications to match fields from one database to another, data extraction, code execution and reviewing for correctness. Once finished, focus should shift to loading the output, which entails storing the transformed data in the appropriate dataset.

Exploratory Data Analysis (EDA)

Exploratory data analysis or EDA involves data set analysis with an objective of identifying patterns, anomalies and relationships amongst your data.

EDA involves three basic steps: data profiling, visualization techniques and the identification of anomalies and other issues within the data set.

- **Data Profiling -** Using statistical summaries, you can gain an understanding of your data distribution, missing values, outliers and other anomalies.

- **Visualization Techniques** - Leverage charts and plots, including scatter plots and histograms to more clearly visualize patterns, trends and relationships amongst data sets.

- **Identify Potential Problem Areas** - Review your data sets to pinpoint inconsistencies, correlations and feature importance that may require special attention down the road.

Feature Selection

Feature selection has a significant impact on user experience and the overall functionality of your platform. Dimensionality reduction is key. This involves identifying relevant features and then examining those features with a goal of improving interpretability and reducing overfitting.

To achieve this, you can call upon several techniques to select the most important features, such as correlation analysis, PCA or lasso regression. This approach includes:

- **Filter methods;**
- **Wrapper methods; and**
- **Embed methods.**

Additionally, a well-thought-out feature selection process can improve model performance by accelerating speeds, increasing accuracy and making for an all-around robust model. Careful feature selection allows you to achieve a degree of optimization that's essential for efficiency and optimization.

Data Splitting

The data model creation process typically involves three sets of data: a training set, a validation set and test sets — all of which are essential for the model development process.

- **The training data set** features examples that are used to fit the parameters of a model. Once complete, this data set is used to train the model using one of several supervised learning methods.

- **The validation data set** (also known as the "dev set") is composed of examples that are used to refine hyperparameters — in essence, the architecture — of a classifier. The validation set helps you to avoid overfitting.

- **The test set** is a critical set of data, which exists independently of the training set, although both follow the same probability distribution. You can test to see whether the model fits well to both the data set and the training set; if so, this suggests that you have a case of minimal overfitting. If the model has a better fit to the training set versus the test set, this is suggestive of a potentially problematic case of overfitting.

In order to avoid overfitting, when any classification parameter needs to be adjusted, it is necessary to have a validation data set.

There are a few additional points that you must consider as well during the data splitting phase of your project.

- **You must also consider the issue of data leakage.** It is essential to split data early in the model development process to ensure that you avoid issues such as overfitting and data leakage between the training set and the test set.

- **Stratified splitting is another key consideration.** With a stratified data-splitting strategy, you can maintain the proper proportions for each class in the training data set and test data set, ultimately ensuring that the proportions align with those that were present in your original dataset. Stratification can be essential for achieving and maintaining class balance, particularly in the case of classification tasks.

- **With cross-validation techniques,** you have the ability to implement methods such as K-fold cross-validation and stratified K-fold validation, which gives you the ability to split data into multiple training sets and validation sets, thereby reducing bias and minimizing variance in model performance estimates.

- **Time-series data splitting** involves the use of techniques such as rolling windows or expanding windows to split time-series data, ultimately ensuring that temporal order is preserved. This approach also allows you to avoid lookahead bias by only using past data to predict future values.

Data Augmentation

There are a number of scenarios where data augmentation can be advantageous. Data augmentation refers to the practice of taking existing data sets and using them to create new samples that serve to improve machine learning models. Data augmentation gives you the ability to "artificially" increase a data set's size and diversity, helping you to avoid overfitting, while simultaneously improving model accuracy and even slashing operational costs.

Synthetic Data Generation

Data augmentation brings the biggest benefit in cases where you have a limited amount of data, such as the case of certain business verticals (e.g. healthcare) or in

cases where you're developing a platform that involves computer vision or NLP technology.

You may see a significant increase in the size of your data pool if you opt to leverage data augmentation techniques, which can include rotating / flipping images and adding noise to text / time series data. The more data you have for machine learning model training, the greater the ROI you'll typically see.

Amplifying Data With GANs

When actual data is limited, you can call upon Generative Adversarial Network (GAN) technology. This data augmentation approach is used to generate synthetic data that closely resembles real data, making it far easier to train a machine learning model when authentic data is scarce. GANs generate realistic artificial data points that reflect the underlying patterns, trends and data distributions in the authentic data set.

Verifying Data Integrity

When working with synthetic data, it's important that you take the time to examine these data sets to verify that they reflect the figures you would see in legitimate data sets. Augmented data must be representative of the patterns and trends that are observed in legitimate data. This step for verifying data integrity is crucial, as it ensures that your data sets are suitable for use in the machine learning model training process.

Data Pipelines

By creating data pipelines, you'll establish efficient and reproducible workflows that allow you to make the most of your data. With a data pipeline, you'll develop modular and reusable data preparation processes, with steps that ensure a standardized, scalable and consistent workflow.

Automation and Scalability

In fact, scalability is essential for successful long term data management. You must be sure that your data pipelines expand as your organization grows over time. The systems and processes that you implement today should remain viable months and years into the future.

Fortunately, it's now possible to automate the tasks and workflows that are associated with a company's data preparation processes. This allows for improved

efficiency and scalability. Tools such as Apache Airflow and Luigi are available for managing and automating data processing tasks.

Other data pipeline management recommendations include the following.

- **Version Control for Improved Consistency -** By implementing version control for data sets, you ensure synchronization between training and production environments. This allows you to maintain data integrity too.

- **Monitoring and Maintenance** - With data pipeline monitoring systems, you can detect anomalies and glitches before an issue dramatically skews your data sets. Deploy monitoring to track performance and establish regular maintenance procedures to keep your data in tip top shape.

- **Documentation -** Maintain comprehensive documentation for your data preparation workflows, versioning and changes to ensure clarity, transparency and reproducibility.

Data Prep Drives The Most Innovative, Lucrative AI Projects

Data preparation is essential for the success of your AI development project — or any other development project, for that matter.

Contact the team at 7T today to get started with your AI development project.

https://7T.co

Houston Regional Office

1334 Brittmoore Road
Suite D
Houston, Texas 77043
+1 (832) 632-4869

Dallas Headquarters

16803 Dallas Parkway
Suite 300
Addison, Texas 75001
+1 (214) 299-5100

Charlotte Regional Office

2115 Rexford Road
Suite #570
Charlotte, NC 28211
+1 (980) 350-5100