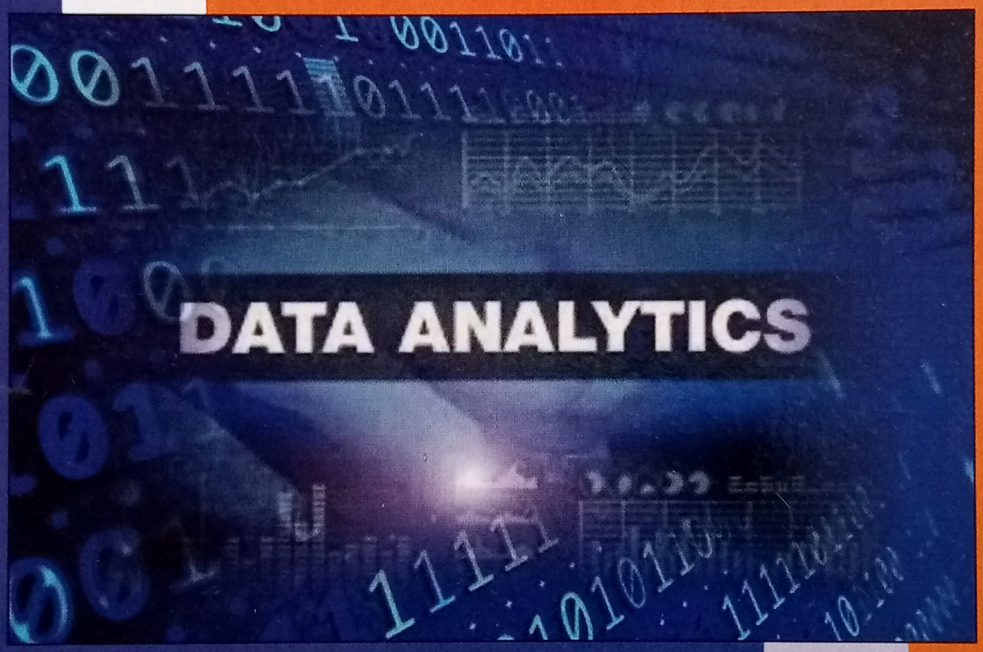
T. Y. B. Sc.
COMPUTER SCIENCE
SEMESTER-VI

NEW SYLLABUS
CBCS PATTERN

DATA ANALYTICS

Dr. Ms. MANISHA BHARAMBE
Dr. Mrs. HARSHA PATIL



NIRALI PRAKASHAN
ADVANCEMENT OF KNOWLEDGE

Contents ...

■■■

CHAPTER 1

Introduction to Data Analytics

Objectives...

- To understand Concept of Data Analytics
- To learn Types of Data Analytics
- To study different Types of Data Analytics

1.0 INTRODUCTION

- An important phase of technological innovation associated with the rise and rapid development of computer technology came into existence only a few decades ago.
- The technological innovation brought about a revolution in the way people work, first in the field of science and then in many others, from technology to business, as well as in day-to-day life.
- In today's data driven world the massive amount of data collected/generated/ produced at remarkable speed and high volume at every day. Data allows to makes better predictions about the future.
- For processing and analyzing need of this massive/huge amount of a new discipline is formed known as data science. The objective/goal of data science is to extract information from data sources.
- Data science is a collection of techniques used to extract value from data. Data science has become an essential tool for any organization that collects stores and processes data as part of its operations.
- Data science is the task of scrutinizing and processing raw data to reach a meaningful conclusion. Data science techniques rely on finding useful patterns, connections and relationships within data.
- Data science applies an ever-changing and vast collection of techniques and technology from mathematics, statistics, Machine Learning (ML) and Artificial Intelligence (AI) to decompose complex problems into smaller tasks to deliver insight and knowledge.
- Analytics is the systematic computational analysis of data. Analytics is the discovery, interpretation, and communication of meaningful patterns in data.

- Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.
- Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

1.1 CONCEPT OF DATA ANALYTICS

- Advancement in data science has created opportunities to sort, manage and analyze large or massive amounts of data more effectively and efficiently.
- Data science is closely related to the fields of data mining and machine learning, but it is broader in scope. Today, data science drives decision making in nearly all parts of modern societies.
- The term data comprises facts, observations and raw information. Data itself have little meaning if it is not processed. The processed data in meaningful form known as information.
- Analytics is used for the discovery, interpretation, and communication of meaningful patterns and/or insights in data. The term analytics is used to refer to any data-driven decision-making.
- Data analytics may analyze many varieties of data to provide views into patterns and insights that are not humanly possible.
- Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions from that information.
- Data analytics is used in many industries to allow companies and organizations to make better business decisions, and in the sciences to verify or disprove existing models or theories.

1.1.1 Definition of Data Analytics

- Data and information are increasing rapidly; the growth rate of the information is so high that the information available to us in the near future is going to unpredictable.
- So, there is need of technique like data analytics which operates at high-speed and efficiently on huge/massive amount data and helps organizations for making better decisions.
- Data analytics is defined as, a science of extracting meaningful, valuable information from raw data.
- Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.
- The goal of data analytics is to get actionable insights from raw data resulting better decisions.

- Data analytics and all associated strategies and techniques are essential when it comes to identifying different patterns, finding anomalies and relationships in large chunks/set of data and making the data or information collected more meaningful and more understandable.

1.1.2 Roles in Data Analytics

- Various roles in data analytics are explained below:

1. **Data Analyst:** Data analyst is an individual, who performs mining of huge amount of data, models the data, looks for patterns, relationship, trends and so on. He/she comes up with visualization and reporting for analyzing the data for decision making and problem-solving process. The main role of a data analyst is to extract data and interpret the information attained from the data for analyzing the outcome of a given problem.
2. **Data Scientist:** A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc. Data scientists mainly deal with large and complex data that can be of high dimension, and carry out appropriate machine learning and visualization tools to convert the complex data into easily interpretable meaningful information. The primary task of a data scientist is to use machine learning and deep learning-based techniques to make an in-depth analysis of input data.
3. **Data Architect:** They are provides the support of various tools and platforms that are required by data engineers to carry out various tests with precision. Data architects should he well equipped with knowledge of data modeling and data warehousing. The main task of data architects is to design and implement database systems, data models, and components of data architecture.
4. **Data Engineer:** A data engineer works with massive amount of data and responsible for building and maintaining the data architecture of a data science project. Data engineer also works for the creation of data set processes used in modeling, mining, acquisition and verification. Data engineers have a demanding role in data analytics as they help in assuring that data are made available in a form that can be easily used for analysis and interpretation.
5. **Analytics Manager:** They are involved in the overall management of the various data analytics operations as discussed in this section. For each of the stakeholders of data analytics that have been mentioned in this section, the analytics manager deals with the team leader of each group and monitors and manages the work of each team.

1.1.3 Lifecycle of Data Analytics

- The data analytics lifecycle is a process that consists of six basic stages/phases (da discovery, data preparation, model planning, model building, communication resul and operationalize) as shown in Fig. 1.1 that define how information is create gathered, processed used and analyzed for organizational goals.
- Fig. 1.1 shows the six phases of the data analytics lifecycle that is followed one pha after another to complete one cycle.
- The lifecycle of the data analytics provides a framework for the best performances each phase from the creation of the project until its completion.
- Data Analytics (DA) is the science of examining raw data with the purpose of drawin conclusions about the information.

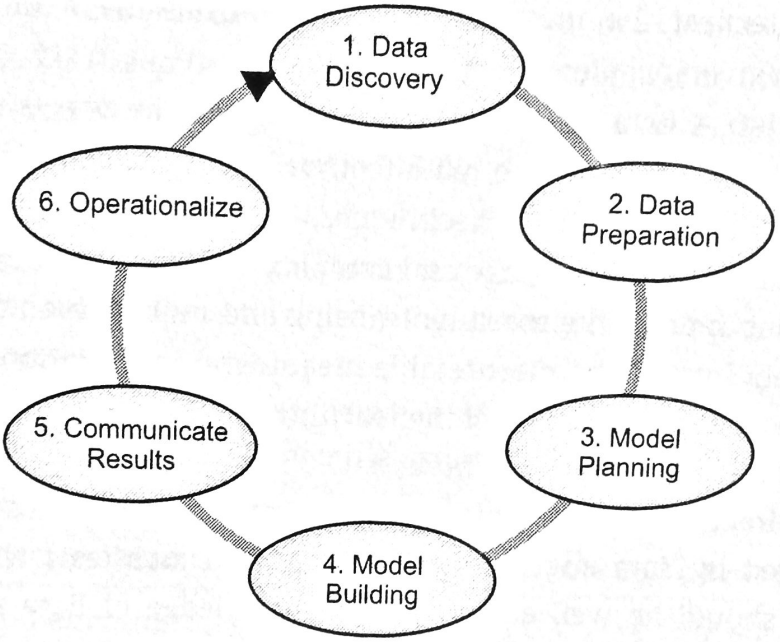


Fig. 1.1: Lifecycle of Data Analytics

- Various phases in data analytics lifecycle are explained below:

Phase 1 - Data Discovery:

- Data discovery is the 1st phase to set project's objectives and find ways to achieve a complete data analytics lifecycle.
- Data discovery phase defining the purpose of data and how to achieve it by the end of the data analytics lifecycle.
- Data discovery phase consists of identifying critical objectives a business is trying to discover by mapping out the data.

Phase 2 - Data Preparation:

- In the 2nd phase after the data discovery phase, data is prepared by transforming it from a legacy system into a data analytics form by using the sandbox platform, (a scalable platform commonly used by the data scientists for data preprocessing).

- Data preparation phase of the data analytics lifecycle involves data preparation, which includes the steps to explore, preprocess and condition data prior to modeling and analysis.
- The data preparation and processing phase involves collecting, processing and conditioning data before moving to the model building process.
- An analytics sandbox is a platform that allows us to store and process large amounts of data.
- Data are loaded in the sandbox in three ways namely, ETL (Extract, Transform and Load), ELT (Extract, Load, and Transform) and ETLT.

Phase 3 - Model Planning:

- The 3rd phase of the lifecycle is model planning, where the data analytics team members makes proper planning of the methods to be adapted and the various workflow to be followed during the next phase of model building.
- Model planning is a phase where the data analytics team members have to analyze the quality of data and find a suitable model for the project.

Phase 4 - Model Building:

- In this phase the team works on developing datasets for training and testing as well as for production purposes.
- This phase is based on the planning made in the previous phase, the execution of the model is carried out by the team.
- Model building is the process where team has to deploy the planned model in a real-time environment. It allows analysts to solidify their decision-making process by gain in-depth analytical information.
- The environment needed for the execution of the model is decided and prepared so that if a more robust environment is required, it is accordingly applied.

Phase 5 - Communicate Results:

- The 5th phase of the life cycle of data analytics checks the results of the project to find whether it is a success or failure.
- The result is scrutinized by the entire team along with its stakeholders to draw inferences on the key findings and summarize the entire work done.
- In communicate results phase, the business/organizational values are quantified and an elaborate narrative on the key findings is prepared.

Phase 6 - Operationalize:

- In 6th phase, the team delivers final reports is prepared by the team along with the briefings, source code and related technical documents.
- Operationalize phase also involves running the pilot project to implement the model and test it in a real-time environment.

- As data analytics help build models that lead to better decision making, it, in turn, adds values to individuals, customers, business sectors and other organizations.
- As soon the team prepares a detailed report including the key findings, documents, and briefings, the data analytics life cycle almost comes close to the end.
- The next step remains the measure the effectiveness of the analysis before submitting the final reports to the stakeholders.

1.1.4 Data Analytics Framework

- In data analytics, the framework allows to move through data analysis in an organized/structured way.
- Data analytics provides us with a process to follow as we scrutinize the data to identify and solve problems.
- Data analytics is the framework deals with technical aspects of managing data and analytics tools. It answers the following questions:
  1. What are the infrastructure requirements today and in 5-10 years?
  2. Should we build an on premise cloud infrastructure or store data in an off premise private virtual cloud?
  3. What are the infrastructure components for data storage and archiving?
  4. Which systems of record will be supported and designated as analytics platforms?
  5. What analytics tools will we support and what will our analytics tools library consist of?
  6. What technologies and vendor solutions will be supported as enterprise analytics systems to provide infrastructure and analytics tools?
  7. What is the analytics capability roadmap?
  8. What solutions do we intend to deploy in the next five years and in what priority?
- Fig. 1.2 shows the four layer framework of data analytics consists of a data management layer an analytics engine layer and a presentation layer.
- The four layers in data analytics framework is explained below:
  1. **Data Connection Layer:** In this layer, data analysts set up data ingestion pipelines and data connectors to access data. They might apply methods to identify metadata (data about data) in all source data repositories. Building this layer starts with making an inventory of where the data is created and stored. The data analysts might implement Extract, Transfer and Load (ETL) software tools to extract data from their source. Other data exchange standards such as X.12 might be used to transfer data to the data management layer. In number of architectures, the enterprise data warehouse may be connected to data sources through data gateways, data harvesters and connectors using APIs. Products offered by Amazon AWS, Microsoft Data Factory and Talend or similar systems are used as data connector tools.

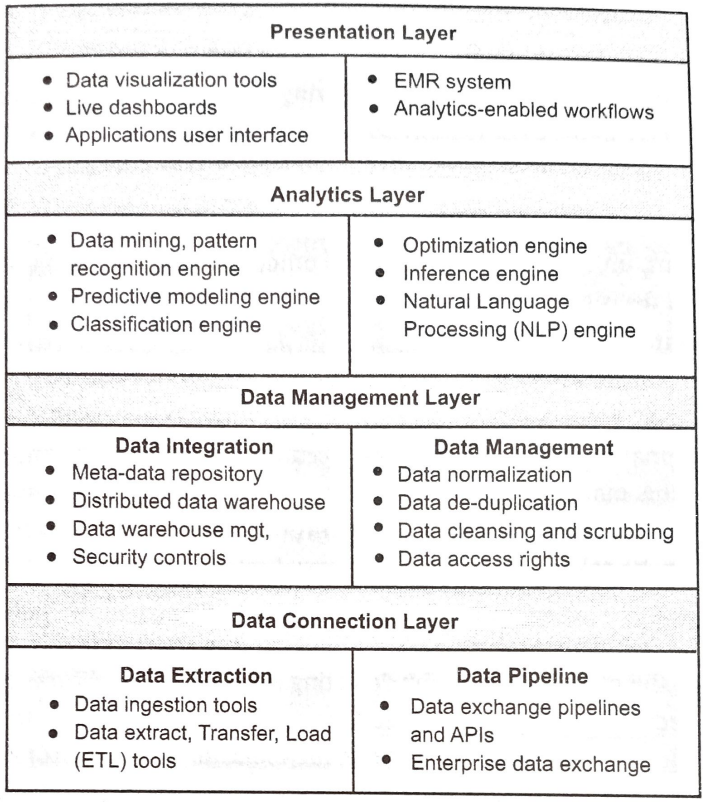


Fig. 1.2: Four Layer Framework of Data Analytics

2. **Data Management Layer:** Once, the data has been extracted, data scientists must perform a number of functions that are grouped under the data management layer. The data may need to be normalized and stored in certain database architectures to improve data query and access by the analytics layer. We'll cover taxonomies of database tools including SQL, NoSQL, Hadoop, Spark and other architecture in the upcoming sections.
3. **Analytics Layer:** In analytics layer, a data scientist uses a number of engines to implement the analytical functions. Depending on the task at hand, a data scientist may use one or multiple engines to build an analytics application. A more complete layer would include engines for optimization, machine learning, natural language processing, predictive modeling, pattern recognition, classification, inferencing and semantic analysis.
4. **Presentation Layer:** The presentation layer includes tools for building dashboards, applications and user-facing applications that display the results of analytics engines. Data scientists often mash up several data visualization widgets,

web parts and dashboards (sometimes called Mash boards) on the screen to display the results using info-graphic reports. These dashboards are active and display data dynamically as the underlying analytics models continuously update the results for dashboards.

1.1.5 Advantages and Disadvantages of Data Analytics

- Data analytics represents the process of examining massive amount of data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions.

Advantages of Data Analytics:

1. **Improving Efficiency:** Data analytics can help analyze large amounts of data quickly and display it in a formulated manner to help achieve specific organizational goals. It encourages a culture of efficiency and teamwork by allowing the managers to share the insights from the analytics results to the employees. The improvement areas within an organization become evident and actions can be taken to increase the overall efficiency of the workplace thereby increasing productivity.
2. **Improving Quality of Products and Services:** Data analytics can help with enhancing the user experience by detecting and correcting errors or avoiding non-value-added tasks. For example, self-learning systems can use data to understand the way customers are interacting with the tools and make appropriate changes to improve user experience. In addition, data analytics can help with automated data cleansing and improving the quality of data and consecutively benefiting both customers and organizations.
3. **Witnessing the Opportunities:** The changing nature of technology the organizations want to keep pace with the latest trends. Here, Data Analytics offers refined sets of data that can help in observing the opportunities to avail.
4. **Helps an Organization make Better Decisions:** Data analytics can help with transforming the data that is available into valuable information for executives so that better decisions can be made.

Disadvantages of Data Analytics:

1. **Low Quality of Data:** One of the biggest limitations of data analytics is lack of access to quality data. It is possible that organizations already have access to a lot of data, but the question is do they have the right data that they need?
2. **Privacy Concerns:** Sometimes, data collection might breach/violate the privacy of the customers as their information such as purchases, online transactions and subscriptions are available to organizations whose services they are using.

1.2 DATA ANALYSIS vs DATA ANALYTICS

- The terms data analysis and data analytics are often used interchangeably and could be confusing.
- Data analytics is a broader term and includes data analysis as necessary subcomponent. Analytics defines the science behind the analysis.
- The science means understanding the cognitive processes an analyst uses to understand problems and explore data in meaningful ways.
- Data analysis is a process that refers to hands-on data exploration and evaluation. Data analysis looks backwards, providing marketers with a historical view of what has happened. Data analytics, on the other hand, models the future or predicts a result.
- The purpose of data analysis is to extract information that is not easily deducible but that, when understood, leads to the possibility of carrying out studies on the mechanisms of the systems that have produced them, thus allowing forecasting possible responses of these systems and their evolution in time.
- Analytics makes extensive use of mathematics and statistics and the use of descriptive techniques and predictive models to gain valuable knowledge from data.
- These insights from data are used to recommend action or to guide decision-making in a business context. Thus, analytics is not so much concerned with individual analysis or analysis steps, but with the entire methodology.
- Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not and what was expected from a product or service.
- On other hand, Data analytics helps organization in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies.
- Data analysis helps to finding or extracting useful information for decision making. Data analytics helps in business growth by reducing risks, costs, and making the right decisions.
- Data Analytics is a wide area involving handling data with a lot of necessary tools to produce helpful decisions with useful predictions for a better output.
- While, Data analysis is actually a subset of data analytics which helps us to understand the data by questioning and to collect useful insights from the information already available.
- Data analytics is the process of exploring the data from the past to make appropriate decisions in the future by using valuable insights whereas, Data analysis helps in understanding the data and provides required insights from the past to understand what happened so far.

- Following table compares data analysis and data analytics:

Sr. No.	Data Analysis	Data Analytics
1.	The process of extracting information from raw data is called as data analysis.	The process of extracting meaningful valuable insights from raw data called as data analytics.
2.	Data analysis is a process involving the collection, manipulation and examination of data for getting insight from data.	Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed organizational decisions.
3.	Data analysis is a process of studying, refining, transforming, and training of the past data to gain useful information, suggest conclusions and make decisions.	Data analytics is the process of remodeling past data into actions through analysis and insights to help in organizational decision making and problem-solving.
4.	Data analysis looks backwards, with a historical view of what has happened.	Data analytics models the future or predicts a result.
5.	Data analysis is a subset of data analytics, which takes multiple data analysis processes to focus on why an event happened and what may happen in the future based on the previous data.	Data analytics is a multidisciplinary field with extensive use of computer skills, mathematics, statistics, the use of descriptive techniques and predictive models to gain valuable knowledge from data through analytics.
6.	Data analysis also makes decisions but less good than data analytics.	Data analytics is utilizing data, machine learning, statistical analysis and computer-based models to get better insight and make better decisions from the data.
7.	Data analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.	Data analytics help uncover the patterns from raw data and derive valuable insights from it.
8.	Data analysis is subset of data analytics.	Data analytics uses data analysis as subcomponent.
9.	Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc.	Tools used in data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.

1.3 TYPES OF DATA ANALYTICS

- Organizations from almost every sector are generating a large volume of data on a regular basis.
- Merely collecting large amounts of data will not serve any purpose and cannot be used directly for the profit of the company/organization.
- Organizations can extract very useful information from this data which can further support complex decision making hence, there is a need for data analytics.
- The art and science of refining data to fetch useful insight which further helps in decision making is known as Analytics.
- There are four types of data analytics as shown in Fig. 1.3 as explained below:
  1. **Descriptive Analytics:** What happened?
  2. **Diagnostic Analytics:** Why did it happen?
  3. **Predictive Analytics:** What will it happen?
  4. **Prescriptive Analytics:** How can we make it happen?

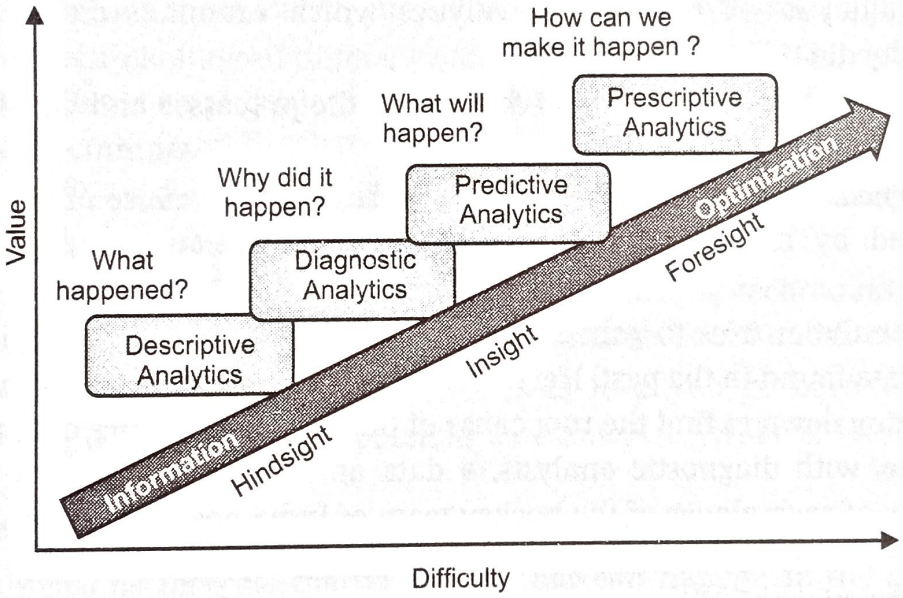


Fig. 1.3: Types of Data Analytics

1.3.1 Descriptive Analytics

- Descriptive analytics examines the raw data or content to answer question, what happened?, by analyzing valuable information found from the available past (historical) data.
- The goal of descriptive analytics is to provide insights into the past leading to the present, using descriptive statistics, interactive explorations of the data, and data mining.

- Descriptive analytics enables learning from the past and assessing how the past might influence future outcomes.
- Descriptive analytics is valuable as it enables associations to gain from past practices and helps them in seeing how they may impact future results.
- Descriptive analytics looks at data and analyzes past events for insight as to how to approach the future.
- It looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure.

Examples:

1. An organizations' records give a past review of their financials, operations, customers and stakeholders, sales and so on.
2. Using descriptive analysis, a data analyst will be able to generate the statistical results of the performance of the hockey players of team India. For generating such results, the data may need to be integrated from multiple data sources to gain meaningful insights through statistical analysis.

1.3.2 Diagnostic Analytics

- Diagnostic analytics is a form of analytics which examines data to answer the question, why did it happen?.
- It is kind of root cause analysis that focuses on the processes and causes, key factors and unseen patterns.
- The goal/objective of diagnostic analytics is to find the root cause of issues. It can be accomplished by techniques like data discovery, correlations, data mining and drill-down.
- Diagnostic analytics tries to gain a deeper understanding of the reasons behind the pattern of data found in the past. Here, business/organizational intelligence comes into play by digging down to find the root cause of the pattern or nature of data obtained.
- For example, with diagnostic analysis, a data analyst will be able to find why the performance of each player of the hockey team of India has risen (or degraded) in the recent past nine months.
- The main function of diagnostic analytics is to identify anomalies, drill into the analytics and determine the casual relationships.

Examples:

1. Some form of social media marketing campaign where the user is interested in retrieving the number of likes or reviews. Diagnostic analytics can help to filter out thousands of likes and reviews into a single view to see the progress of the campaign.
2. Drop in website traffic of an organization can lead to a decrease in the sales and thereby revenue will also be reduced. In this case, diagnostic analytics finds the

root cause initially, such as traffic has been reduced and from there, it will fine-tune the problem after finding the reasons for the downside in website traffic such as Software Engine Optimization (SEO), social marketing, email marketing and any other factors, which are not enabling the website to reach many people.

1.3.3 Predictive Analytics

- Predictive analysis, as the name suggests, deals with prediction of future based on the available current and past data.
- A predictive analysis uses past data to create a model that answer the question, what will happen?
- Prediction-based on historical data, build models and use them to forecast a future value. For example, demand for a particular package around holiday season.
- Predictive analytics is important to analyze the current/present data and make use of it to predict a solution for the future. For these future predictions in data analytics the predictive analytics is used.
- Predictive analytics makes predictions about future outcomes/result using historical data combined with statistical modeling, data mining techniques and machine learning.
- Organizations/firms employ predictive analytics to find patterns in this data to identify risks and opportunities.
- Using predictive analytics, users can prepare plans and implement corrective actions in a proactive manner in advance of the occurrence of an event.
- Predictive analytics is the use of data, machine learning techniques, and statistical algorithms to determine the likelihood of future results based on historical data.
- The primary goal of predictive analytics is to help you go beyond just what has happened and provide the best possible assessment of what is likely to happen in future.
- Predictive analytics extracts the information from the available datasets and this extraction helps us forecast the possibility that can happen in the future with risk analysis and mitigation.
- It is not guaranteed that all the predicted data can produce exact results; there may be a slight variation between the predicted and the future values.
- Based on the past events, a predictive analytics model forecasts what is likely to happen in future.
- Predictive analytics is critical to knowing about future events well in advance and implementing corrective actions.
- Predictive analysis uses techniques to include machine learning, statistical modeling and data mining.

2.	Descriptive analytics	What has happened? How many, when and where?	Canned reports. Ad hoc reports.	It describes the events that have occurs already in the past.
3.	Diagnostics analytics	Why did it happen? Where must we see?	Query and Drilldowns. Discovery alerts.	It justifies the reason for the occurrences of those events in the firm/ organization.
4.	Prescriptive analytics	What should we do about this? What will happen if we use this?	Optimization. Random testing.	It suggests the solutions to overcome the past events.

1.3.5 Exploratory Analytics

- Exploratory Data Analysis (EDA) is the most important aspect to any data analysis. Exploratory data analytics attempts to find hidden, unseen or previously unknown relationships.
- The EDA techniques are used to interactively discover and visualize trends, behaviors, and relationships in data. They also provide clues as to which variables might be good for building data analytics.
- The goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set.
- Exploratory analytics is an analytical approach that primarily focuses on identifying general patterns in the raw data to identify outliers and features that might not have been anticipated using other analytical types.
- EDA is an approach to analyzing datasets to summarize their main characteristics, often with visual/graphical methods.
- The purpose of exploratory data analysis is to:
  1. Check for missing data and other mistakes.
  2. Gain maximum insight into the data set and its underlying structure.
  3. Check assumptions associated with any model fitting or hypothesis test.
  4. Create a list of outliers or other anomalies.
  5. Find parameter estimates and their associated confidence intervals or margins of error.
- Exploratory data analysis is the process of analyzing and interpreting datasets while summarizing their particular characteristics with the help of data visualization methods.

1.4.1 Concept

- In this section, we will see various ways of thinking about models to help shape the way we build them.

Occam's Razor:

- Occam's razor is a problem-solving principle arguing that simplicity is better than complexity.
- Named after 14^{th} century logician and theologian William of Ockham, this theory has been helping many great thinkers for centuries.
- Occam's razor is the problem solving principle, which states that "entities should not be multiplied beyond necessity", sometimes inaccurately paraphrased as "the simplest explanation is usually the best one."
- In simple words, Occam's razor is the philosophical principle states that, the simplest explanation is the best explanation.
- Occam's notion of simpler generally refers to reducing the number of assumptions employed in developing the model.
- With respect to statistical modeling, Occam's razor tells or speaks to the need to minimize the parameter count of a model.
- Overfitting occurs when a mathematical model tries too hard to achieve accurate performance on its training data.
- It is the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably.
- Overfitting occurs or happens when there are so many parameters that the model can essentially memorize its training set, instead of generalizing appropriately to minimize the effects of error and outliers.
- Overfit models tend to perform extremely well on training data, but much less accurately on independent test data.
- An overfit model is a statistical model. An overfit model contains more parameters than can be justified by the data.
- Invoking Occam's razor requires that we have a meaningful way to evaluate how accurately our models are performing. Simplicity is not an absolute virtue, when it leads to poor performance.
- Deep learning is a powerful technique for building models with millions of parameters. Despite the danger of overfitting, these models perform extremely well on a variety of complex tasks.
- Occam would have been suspicious of such models, but come to accept those that have substantially more predictive power than the alternatives.

- Appreciate the inherent trade-off between accuracy and simplicity. It is almost always possible to improve the performance of any model by kludging-on extra parameters and rules to govern exceptions.
- Complexity has a cost, as explicitly captured in machine learning methods like LASSO/ridge regression. These techniques employ penalty functions to minimize the features used in the model.
- Underfitting occurs when a statistical model cannot adequately capture the underlying structure of the data.
- An under-fitted model is a model where some parameters or terms that would appear in a correctly specified model are missing.
- Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model will tend to have poor predictive performance.

Bias-Variance Trade-Offs:

- The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.
- Bias-variance trade-off is tension between the between model complexity and performance shows up in the statistical notion of the bias-variance trade-off:
  1. **Bias:** It is error from incorrect assumptions built into the model, such as restricting an interpolating function to be linear instead of a higher-order curve.
  2. **Variance:** It is error from sensitivity to fluctuations in the training set. If our training set contains sampling or measurement error, this noise introduces variance into the resulting model.
- Errors of bias produce or generate underfit models and they do not fit the training data as tightly as possible, were they allowed the freedom to do so.
- Underfitting occurs/happens when a statistical model cannot adequately capture the underlying structure of the data.
- Errors of variance result in overfit models (their quest for accuracy causes overfit models to mistake noise for signal and they adjust so well to the training data that noise leads them astray).
- Models that do much better on testing data than training data are overfit models. An underfitted model is a model where some parameters or terms that would appear in a correctly specified model are missing.

Nate Silver's Principles for Effective Modeling:

- Nate R. Silver is perhaps the most prominent public face of data science today. He outlines following principles for effective modeling:

  Principle #1 (Think Probabilistically): Forecasts which make concrete statements are less meaningful than those that are inherently probabilistic. The real world is an

uncertain place, and successful models recognize this uncertainty. There are always a range of possible outcomes that can occur with slight perturbations of reality, and this should be captured in the model.

Principle #2 (Change the Forecast in Response to New Information): Live models are much more interesting than dead ones. A model is live if it is continually updating predictions in response to new information. Fresh information should change the result of any forecast. Scientists should be open to changing opinions in response to new data and built the infrastructure that maintains a live model. Any live model should track and display its predictions over time, so the viewer can guess whether changes accurately reflected the impact of new information.

Principle #3 (Look for Consensus): Data should derive from as many different sources as possible to get the good forecast. Ideally, multiple models should be built, each trying to predict the same thing in different ways. We should have an opinion as to which model is the best, but be concerned when it substantially differs from the herd. Often third parties produce competing forecasts, which you can monitor and compare against.

Principle #4 (Employ Baysian Reasoning): The Bayes' theorem has several interpretations, but perhaps most clearly provides a way to calculate how probabilities change in response to new evidence. When stated as given below, it provides a way to calculate how the probability of event A changes in response to new evidence B.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Applying Bayes' theorem requires a prior probability P(A), the likelihood of event A before knowing the status of a particular event B. This might be the result of running a classifier to predict the status of A from other features or background knowledge about event frequencies in a population. Without a good estimate for this prior, it is very difficult to know how seriously to take the classifier.

1.4.2 Taxonomy of Models

- Mathematical models come in, well, many different models. Part of producing or developing a philosophy of modeling understands the available degrees of freedom in design and implementation.
- In this section, we will study at model types along several different dimensions, reviewing the primary technical issues which arise to distinguish each class.

First-Principle vs. Data-Driven Models:

- First-principle models are based on a belief of how the system under investigation really works. First-principle models might be a theoretical explanation, like Newton's laws of motion.

- By contrast, we prefer mathematical models that are descriptive, meaning these models provide some insight into why they are making their decisions.
- Generally, theory driven models are descriptive because they are explicit implementations of a particular well developed theory.
- If we believe the theory, we have a reason to trust the underlying model, and any resulting predictions. Certain Machine Learning (ML) models prove less opaque than others.
- Linear regression models are descriptive in nature. Because one can see exactly which variables receive the most weight, and measure how much they contribute to the resulting prediction.
- Decision tree models enable us to follow the exact decision path used to make a classification.
- But the unfortunate truth is that black box modeling techniques such as deep learning can be extremely effective.
- Generally, neural network models are completely opaque as to why they do what they do.

Flat vs. Hierarchical Models:

- Interesting problems often exist on several different levels, each of which may require independent sub-models.
- Predicting the future price for a particular stock really should involve sub-models for analyzing such separate issues as given below:
  - the general state of the economy,
  - the company's balance sheet, and
  - the performance of other companies in its industrial sector.
- Imposing a hierarchical structure on a model permits it to be built and evaluated in a logical and transparent way, instead of as a black box.
- Certain sub-problems lend themselves to theory-based models, first-principle models, which can then be used as features in a general data driven model.
- Explicitly hierarchical models are descriptive in nature (one can trace a final decision back to the appropriate top-level sub-problem, and report how strongly it contributed to making the observed result).
- The first step to build a hierarchical model is explicitly decomposing the problem into sub-problems. Basically, these represent mechanisms governing the underlying process being modeled.
- Deep learning models in mathematics can be thought of as being both at and hierarchical, at the same time.
- Deep learning models are typically trained on large sets of unwashed data, so there is no explicit definition of sub-problems to guide the sub-process.

- Looked at as a whole, the network does only one thing. But because they are bui from multiple nested layers (the deep in deep learning), these deep learning mode presume that there are complex features there to be learned from the lower lev inputs.

Stochastic vs. Deterministic Models:

- Demanding a single deterministic prediction from a mathematical model can be fool's errand. The world is a complex and critical place of many realities, with even that generally would not unfold in exactly the same way if time could be run ov again.
- Good forecasting models incorporate such thinking and produce probabili distributions over all possible events.
- Stochastic (meaning "randomly determined") modeling techniques that explicitly buil some notion of probability into the model include logistic regression and Monte Carl simulation.
- It is important that the model observe the basic properties of probabilities, including:
  - **Each probability is a value between 0 and 1:** Scores that are not constrained t be in 0 and 1 range do not directly estimate probabilities. The solution is often t put the values through a logit() function to turn them into probabilities in principled way.
  - **Rare events do not have probability zero:** Any event i.e., possible must have greater than zero probability of occurrence. Discounting is a way of evaluating th likelihood of unseen but possible events. Probabilities are a measure of humili about the accuracy of our model and the uncertainty of a complex world. Mode must be honest in what they do and don't know.
  - **That they must sum to 1:** Independently generating values between 0 and 1 do not mean that they together add up to a unit probability, over the full event spac The solution here is to scale these values so that they do, by dividing each by th partition function. Alternately, rethink the model to understand why they didr add up in the first place.

1.4.3 Baseline Models

- To assess the complexity of the task involves building baseline models (the simple reasonable models that produce answers we can compare against).
- More sophisticated models should do better than baseline models, but verifying th they really do and, if so by how much, puts its performance into the proper context.
- Certain forecasting tasks are inherently harder than others. A simple baseline (ye has proven very accurate in predicting whether the sun will rise tomorrow.
- By contrast, we could get rich predicting whether the stock market will go up or dow 51% of the time. Only after we decisively beat our baselines can our models really b deemed effective.

- There are two common tasks for data science models namely, classification and value prediction.

Baseline Models for Classification:

- In classification tasks, we are given a small set of possible labels for any given item, like (man or woman), (spam or not spam) or (car or truck).
- We seek a system that will generate or produce a label accurately describing a particular instance of a person, e-mail or vehicle.
- Representative baseline models for classification include:

1. **Uniform or Random Selection among Labels:** If we have absolutely no prior distribution on the objects, we might as well make an arbitrary selection using the broken watch method. Comparing the stock market prediction model against random coin flips will go a long way to showing how hard the problem is.
2. **The most common Label appearing in the Training Data:** A large training dataset usually provides some notion of a prior distribution on the classes. Selecting the most frequent label is better than selecting them uniformly or randomly. This is the theory behind the sun-will-rise-tomorrow baseline model.
3. **The most Accurate Single-feature Model:** Powerful classification baseline models strive to exploit all the useful features present in a given data set. But it is valuable to know what the best single feature can do. Occam's razor deems the simplest and easiest model to be best. Only when the complicated model beats all single-factor models does it start to be interesting.
4. **Somebody else's Model:** Often we are not the first person to attempt a particular task. Our firm/organization may have a legacy model that we are charged with updating or revising. One of two things can happen/occur when we compare the model against someone else's work: either we beat them or we don't. If we beat them, we now have something worth bragging about. If we don't, it is a chance to learn and improve. Why didn't we win? The fact that we lost gives we certainty that your model can be improved, at least to the level of the other guy's model.
5. **Clairvoyance:** There are circumstances when even the best possible classification baseline model cannot theoretically reach 100% accuracy.

Baseline Models for Value Prediction:

- In baseline value prediction models problems, we are given a collection of feature value pairs (f_i, v_i) to use to train a function F such that $F(v_i) = v_i$.
- Baseline models for value prediction problems follow from similar techniques to what were proposed for classification, as follows:

1. **Mean or Median:** Just ignore the features, so we can always output the consensus value of the target. This proves that, to be quite an informative baseline, because if we cannot substantially beat always guessing the mean, either we have the wrong features or is working on a hopeless task.

- Evaluating a classifier means measuring how accurately our predicted labels match the gold standard labels in the evaluation set.
- For the common case of two distinct labels or classes (binary classification), we typically call the smaller and more interesting of the two classes as positive and the larger/other class as negative.
- In a spam classification problem, the spam would typically be positive and the ham (non-spam) would be negative.
- This labeling aims to ensure that identifying the positives is at least as hard as identifying the negatives, although often the test instances are selected so that the classes are of equal cardinality.
- There are four possible results of what the classification model could do on any given instance, which defines the confusion matrix or contingency table shown in Fig. 1.4.
- A confusion matrix contains information about actual and predicted classifications done by a classifier. Performance of such systems is commonly evaluated using the data in the matrix.
- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
- The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. A confusion matrix also known as an error matrix.
- A confusion matrix is a technique for summarizing the performance of a classification algorithm.
- A confusion matrix is nothing but a table with two dimensions viz. "Actual" and "Predicted" and furthermore, both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)".

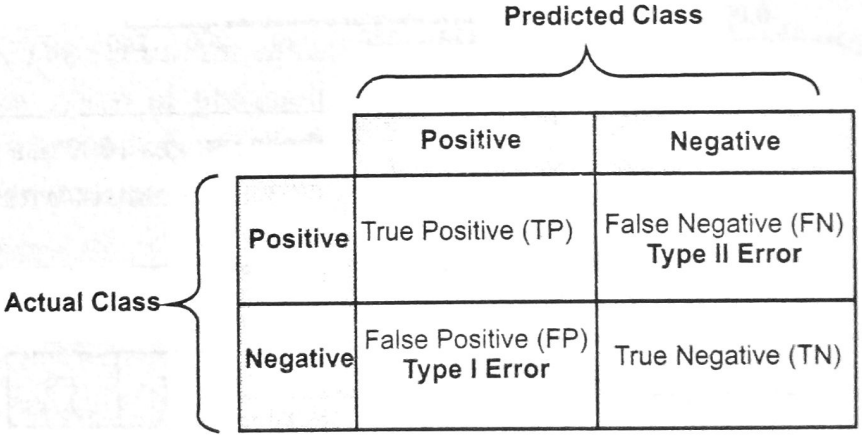


Fig. 1.4: Confusion Matrix for Binary Classifiers, Defining different Classes of Correct and Erroneous Predictions

- The rows in a confusion matrix represent actual class while the columns represent predicted class.

True Positive: We predicted positive and it's true. In the Fig. 1.6, we predicted that 90 images are cat images.

True Negative: We predicted negative and it's true. In the Fig. 1.6, we predicted that 940 images are non cat.

False Positive (Type I Error): We predicted positive and it's false. In the Fig. 1.6, we predicted that 60 images are cat images but actually not.

False Negative (Type II Error): We predicted negative and it's false. In the Fig. 1.6, we predicted that 10 images are non cat but actually yes.

Statistic Measures for Classifier:

1. **Accuracy:** The accuracy of classifier, the ratio of the number of correct predictions over total predictions. We can calculate accuracy by confusion matrix with the help of following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

2. **Precision:** The precision measures the number of positive values predicted by the classifier that are actually positive. We can calculate precision by confusion matrix with the help of following formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall or Sensitivity:** Recall determines the proportion of the positives values that were accurately predicted. Sensitivity or Recall means out of all actual positives, how many did we predict as positive. We can calculate recall by confusion matrix with the help of following formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. **F-score:** The F-score (or sometimes F1-score) is such a combination, returning the harmonic mean of precision and recall. We can calculate recall by confusion matrix with the help of following formula:

$$F = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

- For above example, consider following the confusion matrix:

  [[90 60]
  [10 940]]

  TP = 90, FN = 10, TN = 940 and FP = 60

- For above binary classifier,

  TP + TN = 90 + 940 = 1030 and

  TP + FP + FN + TN = 90 + 60+ 10 + 940 = 1100

Hence, Accuracy = 1030 / 1100 = 0.9364.

- Consider what happens as we sweep our threshold from left to right over these distributions.
- Every time we pass over another example, we either increase the number of true positives (if this example was positive) or false positives (if this example was in fact a negative).
- At the very left, we achieve true/false positive rates of 0%, since the classifier labeled nothing as positive at that cutoff.
- Moving as far to the right as possible, all examples will be labeled positively, and hence both rates become 100%.
- Each threshold in between defines a possible classifier, and the sweep defines a staircase curve in true/false positive rate space taking us from (0%, 0%) to (100%, 100%).
- An ROC curve is the most commonly used way to visualize the performance of a binary classifier and AUC is (arguably) the best way to summarize its performance in a single number.
- The area under the ROC curve (AUC) is often used as a statistic measuring the quality of scoring function defining the classifier.
- The best possible ROC curve has an area of 100% × 100 % → 1, while the monkey's triangle has an area of 1/2. The closer the area is to 1, the better the classification function is.
- The Area Under the Curve (AUC) is another evaluation metric that we can use for classification models.
- The 45 degree line is the baseline for which the AUC is 0.5. The perfect model will have an AUC of 1.0. The closer the AUC to 1.0, the better the predictions.

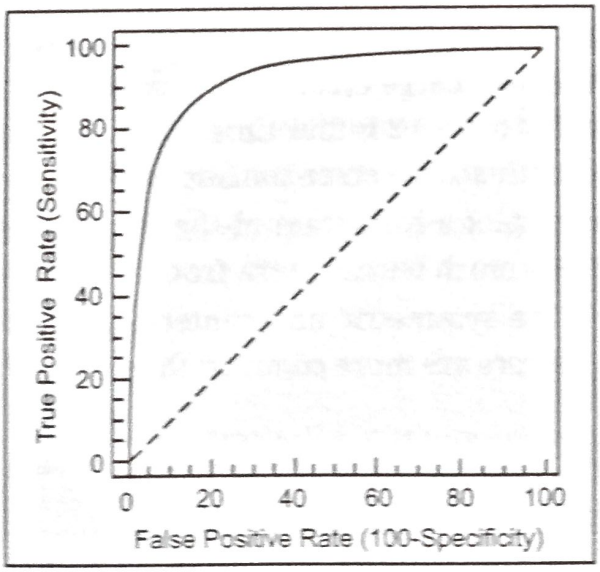


Fig. 1.7: ROC Curve

- Fig. 1.8 shows the absolute error distributions from two models for predicting the year of authorship of documents from their word usage distribution.
- On the left, we see the error distribution for the monkey, randomly guessing a year from 1800 to 2005. What do we see? The error distribution is broad and bad, as we might have expected, but also asymmetric.
- Far more documents produced positive errors than negative ones. Why? The test corpus apparently contained more modern documents than older ones, so is more often positive than negative.

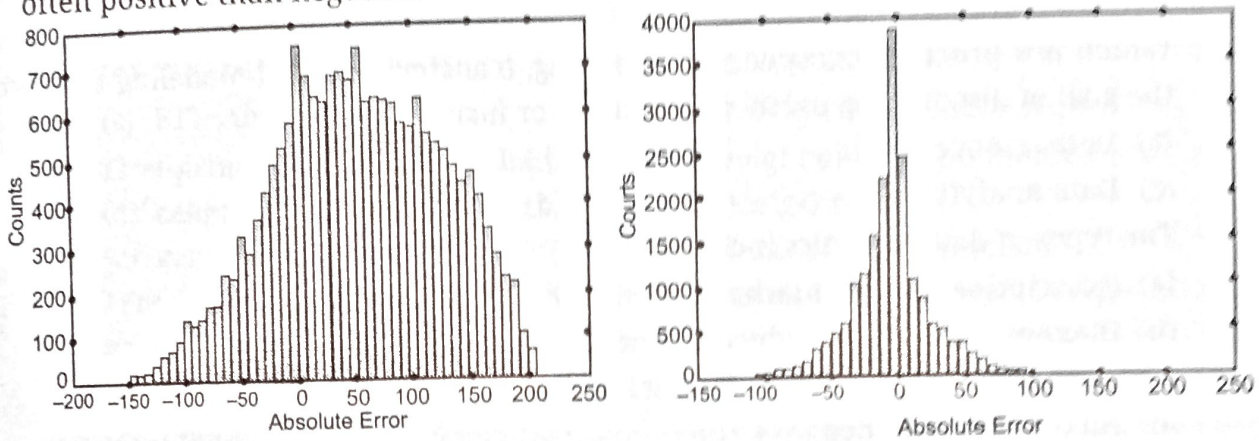


Fig. 1.8: Error Distribution Histograms for Random (Left) and Naive Bayes Classifiers Predicting the Year of Authorship for Documents (Right)

- In contrast, Fig. 1.8 (right) presents the error distribution for our naïve Bayes classifier for document dating. This looks much better: there is a sharp peak around zero and much narrower tails.
- But the longer tail now resides to the left of zero, telling us that we are still calling a distressing number of very old documents modern. We need to examine some of these instances, to figure out why that is the case.
- We need a summary statistic reducing such error distributions to a single number, in order to compare the performance of different value prediction models.
- A commonly-used statistic is Mean Squared Error (MSE), which is computed as follows:

$$\text{MSE}(Y, Y') = \frac{1}{n} \sum_{i=1}^{n} (y'_i - y_i)^2$$

- Because it weighs each term quadratically, outliers have a disproportionate effect. Thus, median squared error might be a more informative statistic for noisy instances.
- Root Mean Squared (RMSD) error is simply the square root of mean squared error:

$$\text{RMSD}(\Theta) = \sqrt{\text{MSE}(Y, Y')}.$$