# Projects Repository

Davut Ayan

2024-07-19

# Contents

# Preface

Hello there! As a devoted explorer navigating the expansive realm of machine learning, I am delighted to present my personal repository—a virtual haven that houses my notes, musings, and sample projects sourced from a diverse array of blogs, books, and practical encounters.

This curated collection serves as a mosaic of insights, with some of the codes and notes thoughtfully extracted from publicly available machine learning blogs. Each project within this repository is a testament to my ongoing quest for understanding, meticulously pieced together from the rich tapestry of the digital knowledge landscape.

Whether you are a fellow enthusiast, a curious mind, or a seasoned practitioner, I extend an invitation to explore the codebase, delve into the concepts, and perhaps find inspiration for your own machine learning journey. This repository is not merely a repository of algorithms and snippets; it is a reflection of my commitment, curiosity, and enthusiasm for the ever-evolving field of machine learning.

I encourage you to engage, share your thoughts, or even collaborate on this journey. Let's celebrate the collaborative spirit of the machine learning community and together, embrace the boundless opportunities that arise from the fusion of code, data, and the collective wisdom of publicly available resources.

Happy exploration!

# Chapter 1

# Projects

DataTab Statistics Tutorials

# Chapter 2

# Machine Learning Fundamentals

## 2.1 definitions

### 2.1.1 Data Science

#### 2.1.1.1 What is data science?

At its core, data science is using data to answer questions. This is a pretty broad definition, and that's because it's a pretty broad field!

Data science can involve:

• Statistics, computer science, mathematics • Data cleaning and formatting • Data visualization An Economist Special Report sums up this mélange of skills well - they state that a data scientist is broadly defined as someone: "who combines the skills of software programmer, statistician and storyteller slash artist to extract the nuggets of gold hidden under mountains of data" And by the end of these courses, hopefully you will feel equipped to do just that!

#### 2.1.1.2 Why do we need data science?

One of the reasons for the rise of data science in recent years is the vast amount of data currently available and being generated. Not only are massive amounts of data being collected about many aspects of the world and our lives, but we simultaneously have the rise of inexpensive computing. This has created the perfect storm in which we have rich data and the tools to analyse it: Rising computer memory capabilities, better processors, more software and now, more data scientists with the skills to put this to use and answer questions using this data! There is a little anecdote that describes the truly exponential growth of data generation we are experiencing. In the third century BC, the Library of

Alexandria was believed to house the sum of human knowledge. Today, there is enough information in the world to give every person alive 320 times as much of it as historians think was stored in Alexandria's entire collection. And that is still growing.

#### 2.1.1.3 What is big data?

It has been so integral to the rise of data science. There are a few qualities that characterize big data. The first is `volume.` As the name implies, big data involves large datasets - and these large datasets are becoming more and more routine. For example, say you had a question about online video - well, YouTube has approximately 300 hours of video uploaded every minute! You would definitely have a lot of data available to you to analyse, but you can see how this might be a difficult problem to wrangle all of that data!

And this brings us to the second quality of big data: `velocity.` Data is being generated and collected faster than ever before. In our YouTube example, new data is coming at you every minute! In a completely different example, say you have a question about shipping times or routes. Well, most transport trucks have real time GPS data available - you could in real time analyse the trucks movements… if you have the tools and skills to do so!

The third quality of big data is `variety.` In the examples I've mentioned so far, you have different types of data available to you. In the YouTube example, you could be analysing video or audio, which is a very unstructured data set, or you could have a database of video lengths, views or comments, which is a much more structured dataset to analyse.

#### 2.1.1.4 1. Descriptive analysis

The goal of descriptive analysis is to describe or summarize a set of data. Whenever you get a new dataset to examine, this is usually the first kind of analysis you will perform. Descriptive analysis will generate simple summaries about the samples and their measurements. You may be familiar with common descriptive statistics: measures of central tendency (eg: mean, median, mode) or measures of variability (eg: range, standard deviations or variance). This type of analysis is aimed at summarizing your sample – not for generalizing the results of the analysis to a larger population or trying to make conclusions. Description of data is separated from making interpretations; generalizations and interpretations require additional statistical steps. Some examples of purely descriptive analysis can be seen in censuses. Here, the government collects a series of measurements on all of the country's citizens, which can then be summarized. Here, you are being shown the age distribution in the US, stratified by sex.

#### 2.1.1.5 2. Exploratory analysis

The goal of exploratory analysis is to examine or explore the data and find relationships that weren't previously known. Exploratory analyses explore how

different measures might be related to each other but do not confirm that relationship as causitive. You've probably heard the phrase "Correlation does not imply causation" and exploratory analyses lie at the root of this saying. Just because you observe a relationship between two variables during exploratory analysis, it does not mean that one necessarily causes the other. Because of this, exploratory analyses, while useful for discovering new connections, should not be the final say in answering a question! It can allow you to formulate hypotheses and drive the design of future studies and data collection, but exploratory analysis alone should never be used as the final say on why or how data might be related to each other.

#### 2.1.1.6 3. Inferential analysis

The goal of inferential analyses is to use a relatively small sample of data to infer or say something about the population at large. Inferential analysis is commonly the goal of statistical modelling, where you have a small amount of information to extrapolate and generalize that information to a larger group.

Inferential analysis typically involves using the data you have to estimate that value in the population and then give a measure of your uncertainty about your estimate. Since you are moving from a small amount of data and trying to generalize to a larger population, your ability to accurately infer information about the larger population depends heavily on your sampling scheme - if the data you collect is not from a representative sample of the population, the generalizations you infer won't be accurate for the population.

#### 2.1.1.7 4. Predictive analysis

The goal of predictive analysis is to use current data to make predictions about future data. Essentially, you are using current and historical data to find patterns and predict the likelihood of future outcomes. Like in inferential analysis, your accuracy in predictions is dependent on measuring the right variables. If you aren't measuring the right variables to predict an outcome, your predictions aren't going to be accurate. Additionally, there are many ways to build up prediction models with some being better or worse for specific cases, but in general, having more data and a simple model generally performs well at predicting future outcomes. All this being said, much like in exploratory analysis, just because one variable may predict another, it does not mean that one causes the other; you are just capitalizing on this observed relationship to predict the second variable. A common saying is that prediction is hard, especially about the future. There aren't easy ways to gauge how well you are going to predict an event until that event has come to pass; so evaluating different approaches or models is a challenge.

We spend a lot of time trying to predict things - the upcoming weather, the outcomes of sports events, and in the example we'll explore here, the outcomes of elections. We've previously mentioned Nate Silver of FiveThirtyEight, where

they try and predict the outcomes of U.S. elections (and sports matches, too!). Using historical polling data and trends and current polling, FiveThirtyEight builds models to predict the outcomes in the next US Presidential vote - and has been fairly accurate at doing so! FiveThirtyEight's models accurately predicted the 2008 and 2012 elections and was widely considered an outlier in the 2016 US elections, as it was one of the few models to suggest Donald Trump at having a chance of winning.

#### 2.1.1.8 Causal analysis

The caveat to a lot of the analyses we've looked at so far is that we can only see correlations and can't get at the cause of the relationships we observe. Causal analysis fills that gap; the goal of causal analysis is to see what happens to one variable when we manipulate another variable - looking at the cause and effect of a relationship. Generally, causal analyses are fairly complicated to do with observed data alone; there will always be questions as to whether it is correlation driving your conclusions or that the assumptions underlying your analysis are valid. More often, causal analyses are applied to the results of randomized studies that were designed to identify causation. Causal analysis is often considered the gold standard in data analysis, and is seen frequently in scientific studies where scientists are trying to identify the cause of a phenomenon, but often getting appropriate data for doing a causal analysis is a challenge. One thing to note about causal analysis is that the data is usually analysed in aggregate and observed relationships are usually average effects; so, while on average giving a certain population a drug may alleviate the symptoms of a disease, this causal relationship may not hold true for every single affected individual.

#### 2.1.1.9 Experimental Design

Now that we've looked at the different types of data science questions, we are going to spend some time looking at experimental design concepts. As a data scientist, you are a scientist and as such, need to have the ability to design proper experiments to best answer your data science questions! What does experimental design mean? Experimental design is organizing an experiment so that you have the correct data (and enough of it!) to clearly and effectively answer your data science question. This process involves clearly formulating your question in advance of any data collection, designing the best set-up possible to gather the data to answer your question, identifying problems or sources of error in your design, and only then, collecting the appropriate data. Why should you care?

#### 2.1.1.10 Confounder:

An extraneous variable that may affect the relationship between the dependent and independent variables. In our example, since age affects foot size and literacy is affected by age, if we see any relationship between shoe size and literacy,

the relationship may actually be due to age – age is "confounding" our experimental design! To control for this, we can make sure we also measure the age of each individual so that we can take into account the effects of age on literacy, as well. Another way we could control for age's effect on literacy would be to fix the age of all participants. If everyone we study is the same age, then we have removed the possible effect of age on literacy.

Age is confounding my experimental design! We need to control for this In other experimental design paradigms, a control group may be appropriate. This is when you have a group of experimental subjects that are not manipulated. So if you were studying the effect of a drug on survival, you would have a group that received the drug (treatment) and a group that did not (control). This way, you can compare the effects of the drug in the treatment versus control group.

A control group is a group of subjects that do not receive the treatment, but still have their dependent variables measured In these study designs, there are other strategies we can use to control for confounding effects. One, we can blind the subjects to their assigned treatment group. Sometimes, when a subject knows that they are in the treatment group (eg: receiving the experimental drug), they can feel better, not from the drug itself, but from knowing they are receiving treatment. This is known as the placebo effect. To combat this, often participants are blinded to the treatment group they are in; this is usually achieved by giving the control group a mock treatment (eg: given a sugar pill they are told is the drug). In this way, if the placebo effect is causing a problem with your experiment, both groups should experience it equally.

Blinding your study means that your subjects don't know what group they belong to - all participants receive a "treatment" And this strategy is at the heart of many of these studies; spreading any possible confounding effects equally across the groups being compared. For example, if you think age is a possible confounding effect, making sure that both groups have similar ages and age ranges will help to mitigate any effect age may be having on your dependent variable - the effect of age is equal between your two groups. This "balancing" of confounders is often achieved by randomization. Generally, we don't know what will be a confounder beforehand; to help lessen the risk of accidentally biasing one group to be enriched for a confounder, you can randomly assign individuals to each of your groups. This means that any potential confounding variables should be distributed between each group roughly equally, to help eliminate/reduce systematic errors.

Randomizing subjects to either the control or treatment group is a great strategy to reduce confounders' effects There is one final concept of experimental design that we need to cover in this lesson, and that is replication. Replication is pretty much what it sounds like, repeating an experiment with different experimental subjects. A single experiment's results may have occured by chance; a confounder was unevenly distributed across your groups, there was a systematic error in the data collection, there were some outliers, etc. However, if you can

repeat the experiment and collect a whole new set of data and still come to the same conclusion, your study is much stronger. Also at the heart of replication is that it allows you to measure the variability of your data more accurately, which allows you to better assess whether any differences you see in your data are significant.

#### 2.1.1.11 Beware p-hacking!

One of the many things often reported in experiments is a value called the p-value. This is a value that tells you the probability that the results of your experiment were observed by chance. This is a very important concept in statistics that we won't be covering in depth here, if you want to know more, check out this video explaining more about p-values. What you need to look out for is when you manipulate p-values towards your own end. Often, when your p-value is less than 0.05 (in other words, there is a 5 percent chance that the differences you saw were observed by chance), a result is considered significant. But if you do 20 tests, by chance, you would expect one of the twenty (5%) to be significant. In the age of big data, testing twenty hypotheses is a very easy proposition. And this is where the term p-hacking comes from: This is when you exhaustively search a data set to find patterns and correlations that appear statistically significant by virtue of the sheer number of tests you have performed. These spurious correlations can be reported as significant and if you perform enough tests, you can find a data set and analysis that will show you what you wanted to see. Check out this FiveThirtyEight activity where you can manipulate and filter data and perform a series of tests such that you can get the data to find whatever relationship you want! XKCD mocks this concept in a comic testing the link between jelly beans and acne - clearly there is no link there, but if you test enough jelly bean colours, eventually, one of them will be correlated with acne at p-value $< 0.05$!

#### 2.1.1.12 Data types

- `Continuous variables` are anything measured on a quantitative scale that could be any fractional number. An example would be something like weight measured in kg.
- `Ordinal` data are data that have a fixed, small ($< 100$) number of levels but are ordered. This could be for example survey responses where the choices are: poor, fair, good.
- `Categorical` data are data where there are multiple categories, but they aren't ordered. One example would be sex: male or female. This coding is attractive because it is self-documenting.
- `Missing` data are data that are unobserved and you don't know the mechanism. You should code missing values as NA.
- `Censored` data are data where you know the missingness mechanism on

some level. Common examples are a measurement being below a detection limit or a patient being lost to follow-up. They should also be coded as NA when you don't have the data. But you should also add a new column to your tidy data called, "VariableNameCensored" which should have values of TRUE if censored and FALSE if not.

#### 2.1.1.13 Data scientists in marketing science

Data scientists in marketing science play a crucial role in leveraging data-driven insights to optimize marketing strategies and improve decision-making.

Data scientists in marketing science contribute significantly to the development of targeted, efficient, and impactful marketing campaigns by harnessing the power of data and analytics. Their work helps organizations optimize their marketing spend, enhance customer experiences, and achieve measurable business outcomes.

Here are some key responsibilities and activities that data scientists in marketing science typically engage in:

1. **Data Analysis:**
   - Conducting extensive data analysis to understand customer behavior, market trends, and other relevant metrics.
   - Utilizing statistical methods and machine learning algorithms to extract meaningful patterns and insights from large datasets.
2. Predictive Modeling:
   - Developing and deploying predictive models to forecast future trends, customer behavior, and campaign outcomes.
   - Using machine learning techniques, such as regression analysis, decision trees, and ensemble methods, to build predictive models.
3. Segmentation and Targeting:
   - Creating customer segments based on demographics, behavior, and other relevant factors.
   - Optimizing marketing strategies by targeting specific segments with personalized and relevant content.
4. A/B Testing:
   - Designing and conducting A/B tests to evaluate the effectiveness of different marketing strategies, campaigns, or variations.
   - Analyzing A/B test results to make data-driven recommendations for optimization.
5. **Causal Inference:**
   - Applying advanced causal inference methods to understand the impact of marketing initiatives on customer behavior.
   - Assessing the causal relationships between marketing activities and business outcomes.
6. **Data Visualization:**
   - Creating clear and compelling data visualizations to communicate

complex insights to non-technical stakeholders.
    - Using tools like Tableau, Power BI, or custom scripts to visualize data in a meaningful way.
7. **Optimization Strategies:**
    - Collaborating with marketing teams to develop and optimize marketing strategies based on data insights.
    - Recommending adjustments to campaigns, targeting strategies, and budget allocations for better performance.
8. **Performance Measurement:**
    - Developing key performance indicators (KPIs) and metrics to assess the success of marketing campaigns.
    - Monitoring and evaluating marketing performance against established benchmarks.
9. **Data Management:**
    - Ensuring the quality and integrity of marketing data by cleaning, preprocessing, and validating datasets.
    - Collaborating with data engineers to design and implement data pipelines for efficient data processing.
10. **Communication and Collaboration:**
    - Effectively communicating findings and insights to non-technical stakeholders, including marketing teams and executives.
    - Collaborating with cross-functional teams to align data-driven strategies with overall business objectives.

**Overfitting:**

- Statistical model is too complex
- Too many parameters when compared to the total number of observations.
- Poor Predictive Performance
- Overfitted model overreacts to minor fluctuations in the training data

**Underfitting:**

- Statistical model is too primitive
- Poor Predictive Performance in the training model
- The underfit model under-reacts to even bigger fluctuations.

**Bias:**

Bias is an error introduced in your model because of the oversimplification of a machine learning algorithm. It can lead to underfitting.

In supervised learning, underfitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data.

Also, these kinds of models are very simple to capture the complex patterns in data like Linear and logistic regression.

In supervised learning, overfitting happens when our model captures the noise along with the underlying pattern in data.

It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

**What is variance?**

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data. (over-fitting issue)

**Why is Bias - Variance Trade-off?**

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand, if our model has large number of parameters then it's going to have high variance and low bias. So, we need to find the right/good balance without overfitting and underfitting the data.

**Name three types of biases that can occur during sampling**

In the sampling process, there are three types of biases, which are:

• Selection bias • Under coverage bias • Survivorship bias

**Question:** What do you understand by the Selection Bias? What are its various types?

**Answer:** Selection bias is typically associated with research that doesn't have a random selection of participants. It is a type of error that occurs when a researcher decides who is going to be studied. On some occasions, selection bias is also referred to as the selection effect.

In other words, selection bias is a distortion of statistical analysis that results from the sample collecting method. When selection bias is not taken into account, some conclusions made by a research study might not be accurate. Following are the various types of selection bias:

• Sampling Bias – A systematic error resulting due to a non-random sample of a populace causing certain members of the same to be less likely included than others that results in a biased sample.

• Time Interval – A trial might be ended at an extreme value, usually due to ethical reasons, but the extreme value is most likely to be reached by the variable with the most variance, even though all variables have a similar mean.

• Data – Results when specific data subsets are selected for supporting a conclusion or rejection of bad data arbitrarily.

• Attrition – Caused due to attrition, i.e. loss of participants, discounting trial subjects or tests that didn't run to completion.

4. Discuss Decision Tree algorithm A decision tree is a popular supervised machine learning algorithm. It is mainly used for Regression and Classification. It allows breaks down a dataset into smaller subsets. The decision tree can able to handle both categorical and numerical data.

5. What is Prior probability and likelihood? Prior probability is the proportion of the dependent variable in the data set while the likelihood is the probability of classifying a given observant in the presence of some other variable.

6. Explain Recommender Systems? It is a subclass of information filtering techniques. It helps you to predict the preferences or ratings which users likely to give to a product.

Question: Please explain Recommender Systems along with an application. Answer: Recommender Systems is a subclass of information filtering systems, meant for predicting the preferences or ratings awarded by a user to some product. An application of a recommender system is the product recommendations section in Amazon. This section contains items based on the user's search history and past orders.

7. Name three disadvantages of using a linear model Three disadvantages of the linear model are: • The assumption of linearity of the errors. • You can't use this model for binary or count outcomes • There are plenty of overfitting problems that it can't solve

8. Why do you need to perform resampling?

Resampling is done in below-given cases: • Estimating the accuracy of sample statistics by drawing randomly with replacement from a set of the data point or using as subsets of accessible data • Substituting labels on data points when performing necessary tests • Validating models by using random subsets

9. List out the libraries in Python used for Data Analysis and Scientific Computations. SciPy, Pandas, Matplotlib, NumPy, SciKit, Seaborn
10. What is Power Analysis? The power analysis is an integral part of the experimental design. It helps you to determine the sample size requires to find out the effect of a given size from a cause with a specific level of assurance. It also allows you to deploy a particular probability in a sample size constraint.

11. Explain Collaborative filtering Collaborative filtering used to search for correct patterns by collaborating viewpoints, multiple data sources, and various agents.
12. Discuss 'Naive' in a Naive Bayes algorithm? The Naive Bayes Algorithm model is based on the Bayes Theorem. It describes the probability of an event. It is based on prior knowledge of conditions which might be related to that specific event.
13. What is a Linear Regression? Linear regression is a statistical programming method where the score of a variable 'A' is predicted from the score of a second variable 'B'. B is referred to as the predictor variable and A as the criterion variable.
14. State the difference between the expected value and mean value They are not many differences, but both of these terms are used in different contexts. Mean value is generally referred to when you are discussing a probability distribution whereas expected value is referred to in the context of a random variable.
15. What the aim of conducting A/B Testing? AB testing used to conduct random experiments with two variables, A and B. The goal of this testing method is to find out changes to a web page to maximize or increase the outcome of a strategy.
16. What is Ensemble Learning? The ensemble is a method of combining a diverse set of learners together to improvise on the stability and predictive power of the model. Two types of Ensemble learning methods are:

**Bagging** Bagging method helps you to implement similar learners on small sample populations. It helps you to make nearer predictions. Boosting Boosting is an iterative method which allows you to adjust the weight of an observation depends upon the last classification. Boosting decreases the bias error and helps you to build strong predictive models. 18. Explain Eigenvalue and Eigenvector Eigenvectors are for understanding linear transformations. Data scientist need to calculate the eigenvectors for a covariance matrix or correlation. Eigenvalues are the directions along using specific linear transformation acts by compressing, flipping, or stretching. Question: Please explain Eigenvectors and Eigenvalues. Answer: Eigenvectors help in understanding linear transformations. They are calculated typically for a correlation or covariance matrix in data analysis. In other words, eigenvectors are those directions along which some particular linear transformation acts by compressing, flipping, or stretching. Eigenvalues can be understood either as the strengths of the transformation in the direction of the eigenvectors or the factors by which the compressions happens.

19. Define the term cross-validation Cross-validation is a validation technique for evaluating how the outcomes of statistical analysis will generalize for an Independent dataset. This method is used in backgrounds where the objective is forecast, and one needs to estimate how accurately a model will accomplish. Question: Can you compare the validation set with the test set? Answer: A validation set is part of the training set used for parameter selection as well as for avoiding overfitting of the machine learning model

being developed. On the contrary, a test set is meant for evaluating or testing the performance of a trained machine learning model.

20. Explain the steps for a Data analytics project The following are important steps involved in an analytics project: • Understand the Business problem • Explore the data and study it carefully. • Prepare the data for modeling by finding missing values and transforming variables. • Start running the model and analyze the Big data result. • Validate the model with new data set. • Implement the model and track the result to analyze the performance of the model for a specific period.

Question: What do you mean by cluster sampling and systematic sampling? Answer: When studying the target population spread throughout a wide area becomes difficult and applying simple random sampling becomes ineffective, the technique of cluster sampling is used. A cluster sample is a probability sample, in which each of the sampling units is a collection or cluster of elements. Following the technique of systematic sampling, elements are chosen from an ordered sampling frame. The list is advanced in a circular fashion. This is done in such a way so that once the end of the list is reached, the same is progressed from the start, or top, again.

23. What is a Random Forest? Random forest is a machine learning method which helps you to perform all types of regression and classification tasks. It is also used for treating missing values and outlier values.
24. What is the importance of having a selection bias? Selection Bias occurs when there is no specific randomization achieved while picking individuals or groups or data to be analyzed. It suggests that the given sample does not exactly represent the population which was intended to be analyzed.
25. What is the K-means clustering method? K-means clustering is an important unsupervised learning method. It is the technique of classifying data using a certain set of clusters which is called K clusters. It is deployed for grouping to find out the similarity in the data.
26. Explain the difference between Data Science and Data Analytics Data Scientists need to slice data to extract valuable insights that a data analyst can apply to real-world business scenarios. The main difference between the two is that the data scientists have more technical knowledge then business analyst. Moreover, they don't need an understanding of the business required for data visualization.
27. Explain p-value? When you conduct a hypothesis test in statistics, a p-value allows you to determine the strength of your results. It is a numerical number between 0 and 1. Based on the value it will help you to denote the strength of the specific result.
28. Define the term deep learning Deep Learning is a subtype of machine learning. It is concerned with algorithms inspired by the structure called artificial neural networks (ANN).
29. Explain the method to collect and analyze data to use social media to predict the weather condition. You can collect social media data using

Facebook, twitter, Instagram's API's. For example, for the tweeter, we can construct a feature from each tweet like tweeted date, retweets, list of follower, etc. Then you can use a multivariate time series model to predict the weather condition.

30. When do you need to update the algorithm in Data science? You need to update an algorithm in the following situation: • You want your data model to evolve as data streams using infrastructure • The underlying data source is changing If it is non-stationarity
31. What is Normal Distribution A normal distribution is a set of a continuous variable spread across a normal curve or in the shape of a bell curve. You can consider it as a continuous probability distribution which is useful in statistics. It is useful to analyze the variables and their relationships when we are using the normal distribution curve.
32. Which language is best for text analytics? R or Python? Python will more suitable for text analytics as it consists of a rich library known as pandas. It allows you to use high-level data analysis tools and data structures, while R doesn't offer this feature.
33. Explain the benefits of using statistics by Data Scientists Statistics help Data scientist to get a better idea of customer's expectation. Using the statistic method Data Scientists can get knowledge regarding consumer interest, behavior, engagement, retention, etc. It also helps you to build powerful data models to validate certain inferences and predictions.
34. Name various types of Deep Learning Frameworks • Pytorch • Microsoft Cognitive Toolkit • TensorFlow • Caffe • Chainer • Keras
35. Explain why Data Cleansing is essential and which method you use to maintain clean data Dirty data often leads to the incorrect inside, which can damage the prospect of any organization. For example, if you want to run a targeted marketing campaign. However, our data incorrectly tell you that a specific product will be in-demand with your target audience; the campaign will fail.
36. What is skewed Distribution & uniform distribution? Skewed distribution occurs when if data is distributed on any one side of the plot whereas uniform distribution is identified when the data is spread is equal in the range.
37. When underfitting occurs in a static model? Underfitting occurs when a statistical model or machine learning algorithm not able to capture the underlying trend of the data.
38. Name commonly used algorithms. Four most commonly used algorithm by Data scientist are: • Linear regression • Logistic regression • Random Forest • KNN
39. What is precision? Precision is the most commonly used error metric is n classification mechanism. Its range is from 0 to 1, where 1 represents 100%
40. What is a univariate analysis? An analysis which is applied to none attribute at a time is known as univariate analysis. Boxplot is widely used, univariate model.

41. How do you overcome challenges to your findings? In order, to overcome challenges of my finding one need to encourage discussion, Demonstrate leadership and respecting different options.
42. Explain cluster sampling technique in Data science A cluster sampling method is used when it is challenging to study the target population spread across, and simple random sampling can't be applied.
43. State the difference between a Validation Set and a Test Set A Validation set mostly considered as a part of the training set as it is used for parameter selection which helps you to avoid overfitting of the model being built. While a Test Set is used for testing or evaluating the performance of a trained machine learning model.
44. Explain the term Binomial Probability Formula? "The binomial distribution contains the probabilities of every possible success on N trials for independent events that have a probability of  of occurring."
45. What is a recall? A recall is a ratio of the true positive rate against the actual positive rate. It ranges from 0 to 1.
46. Discuss normal distribution Normal distribution equally distributed as such the mean, median and mode are equal.
47. While working on a data set, how can you select important variables? Explain Following methods of variable selection you can use: • Remove the correlated variables before selecting important variables • Use linear regression and select variables which depend on that p values. • Use Backward, Forward Selection, and Stepwise Selection • Use Xgboost, Random Forest, and plot variable importance chart. • Measure information gain for the given set of features and select top n features accordingly.
48. Is it possible to capture the correlation between continuous and categorical variable? Yes, we can use analysis of covariance technique to capture the association between continuous and categorical variables.
49. Treating a categorical variable as a continuous variable would result in a better predictive model? Yes, the categorical value should be considered as a continuous variable only when the variable is ordinal in nature. So, it is a better predictive model. Question: Recall: What is the proportion of actual positives was identified correctly? TP / (TP + FN) Precision: What is the proportion of positive identifications was actually correct? TP / (TP + FP) Question: A false positive is an incorrect identification of the absence of a condition when it is absent. A false negative is an incorrect identification of the absence of a condition when it is actually present. Question: Please explain the goal of A/B Testing. Answer: A/B Testing is a statistical hypothesis testing meant for a randomized experiment with two variables, A and B. The goal of A/B Testing is to maximize the likelihood of an outcome of some interest by identifying any changes to a webpage. A highly reliable method for finding out the best online marketing and promotional strategies for a business, A/B Testing can be employed for testing everything, ranging from sales emails to search ads and website copy.

Question: Could you explain how to define the number of clusters in a clustering algorithm? Answer: The primary objective of clustering is to group together similar identities in such a way that while entities within a group are similar to each other, the groups remain different from one another. Generally, the Within Sum of Squares is used for explaining the homogeneity within a cluster. For defining the number of clusters in a clustering algorithm, WSS is plotted for a range pertaining to a number of clusters. The resultant graph is known as the Elbow Curve. The Elbow Curve graph contains a point that represents the point post in which there aren't any decrements in the WSS. This is known as the bending point and represents K in K–Means. Although the aforementioned is the widely-used approach, another important approach is the Hierarchical clustering. In this approach, dendrograms are created first and then distinct groups are identified from there.

Question: Please explain Gradient Descent. Answer: The degree of change in the output of a function relating to the changes made to the inputs is known as a gradient. It measures the change in all weights with respect to the change in error. A gradient can also be comprehended as the slope of a function. Gradient Descent refers to escalating down to the bottom of a valley. Simply, consider this something as opposed to climbing up a hill. It is a minimization algorithm meant for minimizing a given activation function.

Question: Please enumerate the various steps involved in an analytics project. Answer: Following are the numerous steps involved in an analytics project: • Understanding the business problem • Exploring the data and familiarizing with the same • Preparing the data for modeling by means of detecting outlier values, transforming variables, treating missing values, et cetera • Running the model and analyzing the result for making appropriate changes or modifications to the model (an iterative step that repeats until the best possible outcome is gained) • Validating the model using a new dataset • Implementing the model and tracking the result for analyzing the performance of the same

Question: What are outlier values and how do you treat them? Answer: Outlier values, or simply outliers, are data points in statistics that don't belong to a certain population. An outlier value is an abnormal observation that is very much different from other values belonging to the set. Identification of outlier values can be done by using univariate or some other graphical analysis method. Few outlier values can be assessed individually but assessing a large set of outlier values require the substitution of the same with either the 99th or the 1st percentile values. There are two popular ways of treating outlier values: 1. To change the value so that it can be brought within a range 2. To simply remove the value Note: - Not all extreme values are outlier values.

21. Discuss Artificial Neural Networks Artificial Neural networks (ANN) are a special set of algorithms that have revolutionized machine learning. It helps you to adapt according to changing input. So the network generates the best possible result without redesigning the output criteria.

22. What is Back Propagation? Back-propagation is the essence of neural net training. It is the method of tuning the weights of a neural net depend upon the error rate obtained in the previous epoch. Proper tuning of the helps you to reduce error rates and to make the model reliable by increasing its generalization.

**Explain Auto-Encoder**

Autoencoders are learning networks. It helps you to transform inputs into outputs with fewer numbers of errors. This means that you will get output to be as close to input as possible.

36. Define Boltzmann Machine Boltzmann machines is a simple learning algorithm. It helps you to discover those features that represent complex regularities in the training data. This algorithm allows you to optimize the weights and the quantity for the given problem.

37. What is reinforcement learning? Reinforcement Learning is a learning mechanism about how to map situations to actions. The end result should help you to increase the binary reward signal. In this method, a learner is not told which action to take but instead must discover which action offers a maximum reward. As this method based on the reward/penalty mechanism.

**Training-Validation-Test**

- We typically train our model but get evaluation metrics on the test data.

From ISLR:

- In general, we do not really care how well the method works on the training data. Rather, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen data. Why is this what we care about? Suppose that we are interested in developing an algorithm to predict a stock's price based on previous stock returns. We can train the method using stock returns from the past 6 months. But we don't really care how well our method predicts last week's stock price. We instead care about how well it will predict tomorrow's price or next month's price.

- We can use patients data to train a statistical learning method to predict risk of diabetes based on clinical measurements. In practice, we want this method to accurately predict diabetes risk for future patients based on their clinical measurements. We are not very interested in whether or not the method accurately predicts diabetes risk for patients used to train the model, since we already know which of those patients have diabetes.

- We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE.

- The problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE. For these methods, the training set MSE can be quite small, but the test MSE is often much larger.

**R-squared**

- In simple linear regression $r^2 = R^2$

From ISLR:

- A number near 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance 2 is high, or both.
- It can still be challenging to determine what is a good R2 value, and in general, this will depend on the application.
- In certain problems in physics, we may know that the data truly comes from a linear model with a small residual error. In this case, we would expect to see an R2 value that is extremely close to 1, and a substantially smaller R2 value might indicate a serious problem with the experiment in which the data were generated.
- On the other hand, in typical applications in biology, psychology, marketing, and other domains, the linear model is at best an extremely rough approximation to the data, and residual errors due to other unmeasured factors are often very large. In this setting, we would expect only a very small proportion of the variance in the response to be explained by the predictor, and an $R^2$ value well below 0.1 might be more realistic!

**F-Test**

Testing whether all of the regression coefficients are zero, i.e. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$Ha$ : at least one $\beta_j$ is non-zero.

## 2.1.2 Bootstrapping

### 2.1.2.1 Jack-knife

• The jackknife is a tool for estimating standard errors and the bias of estimators

• As its name suggests, the jackknife is a small, handy tool; in contrast to the bootstrap, which is then the moral equivalent of a giant workshop full of tools

• Both the jackknife and the bootstrap involve re-sampling data; that is, repeatedly creating new data sets from the original data

The jackknife deletes each observation and calculates an estimate based on the remaining n − 1 of them

• It uses this collection of estimates to do things like estimate the bias and the standard error

• Note that estimating the bias and having a standard error are not needed for things like sample means, which we know are unbiased estimates of population means and what their standard errors are

It has been shown that the jackknife is a linear approximation to the bootstrap

• Generally do not use the jackknife for sample quantiles like the median; as it has been shown to have some poor properties

The bootstrap

• The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistics • For example, how would one derive a confidence interval for the median? • The bootstrap procedure follows from the so called bootstrap principle

Suppose that I have a statistic that estimates some population parameter, but I don't know its sampling distribution • The bootstrap principle suggests using the distribution defined by the data to approximate its sampling distribution • In practice, the bootstrap principle is always carried out using simulation • The general procedure follows by first simulating complete data sets from the observed data with replacement • This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution • Calculate the statistic for each simulated data set • Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error Example • Consider again, the data set of 630 measurements of gray matter volume for workers from a lead manufacturing plant • The median gray matter volume is around 589 cubic centimeters • We want a confidence interval for the median of these measurements • Bootstrap procedure for calculating for the median from a data set of n observations i. Sample n observations with replacement from the observed data resulting in one simulated complete data set ii. Take the median of the simulated data set iii. Repeat these two steps B times, resulting in B simulated medians iv. These medians are approximately draws from the sampling distribution of the median of n observations; therefore we can • Draw a histogram of them • Calculate their standard deviation to estimate the standard error of the median • Take the 2.5th and 97.5th percentiles as a confidence interval for the median

Summary • The bootstrap is non-parametric • However, the theoretical arguments proving the validity of the bootstrap rely on large samples • Better percentile bootstrap confidence intervals correct for bias • There are lots of variations on bootstrap procedures; the book "An Introduction to the Bootstrap" by Efron and Tibshirani is a great place to start for both bootstrap and jackknife information

### 2.1.3 Classification Models: Evaluation

My medium story

Google developers

#### 2.1.3.1 Thresholding

Logistic regression returns a probability. You can use the returned probability "as is" (for example, the probability that the user will click on this ad is 0.00023) or convert the returned probability to a binary value (for example, this email is spam).

A logistic regression model that returns 0.9995 for a particular email message is predicting that it is very likely to be spam. Conversely, another email message with a prediction score of 0.0003 on that same logistic regression model is very likely not spam.

However, what about an email message with a prediction score of 0.6? In order to map a logistic regression value to a binary category, you must define a `classification threshold` (also called the `decision threshold`).

A value above that threshold indicates "spam"; a value below indicates "not spam." It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and are therefore values that you must tune.

Note: "Tuning" a threshold for logistic regression is different from tuning hyperparameters such as learning rate. Part of choosing a threshold is assessing how much you'll suffer for making a mistake. For example, mistakenly labeling a non-spam message as spam is very bad. However, mistakenly labeling a spam message as non-spam is unpleasant, but hardly the end of your job.

#### 2.1.3.2 Confusion Matrix

| | Predicted Positive | Predicted Negative |
|---|---|---|
| **Actual Positive** | TP | FN |
| **Actual Negative** | FP | TN |

**True Positive:** Model predicted positive and it is true.

**True negative:** Model predicted negative and it is true.

**False positive (Type 1 Error):** Model predicted positive but it is false.

**False negative (Type 2 Error):** Model predicted negative and it is true.

**False Positive Rate (FPR):**

The False Positive Rate is the ratio of false positive predictions to the total number of actual negatives. It measures the rate at which the model incorrectly predicts the positive class among the instances that are actually negative.

$FPR = \frac{FP}{TN+FP}$

**True Positive Rate (TPR), Sensitivity, or Recall:**

The True Positive Rate is the ratio of true positive predictions to the total number of actual positives. It measures the ability of the model to correctly predict the positive class among instances that are actually positive.

$Recall(TPR) = \frac{TP}{TP+FN}$

**Accuracy:**

It represents the ratio of correctly predicted instances to the total number of instances. The accuracy metric is suitable for balanced datasets where the classes are evenly distributed. It is calculated using the following formula:

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

Accuracy provides a general sense of how well a model is performing across all classes. It is easy to understand and interpret, making it a commonly used metric, especially when the classes are balanced.

However, accuracy may not be an ideal metric in situations where the class distribution is imbalanced. In imbalanced datasets, where one class significantly outnumbers the other, a high accuracy might be achieved by simply predicting the majority class. In such cases, other metrics like precision, recall, F1 score, or area under the receiver operating characteristic (ROC-AUC) curve may be more informative.

**Precision:**

Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. It represents the accuracy of the positive predictions made by the model.

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

**F1 Measure:**

The F1 score is a metric commonly used in binary classification to provide a balance between precision and recall. It is the harmonic mean of precision and recall, combining both measures into a single value. The F1 score is particularly useful when there is an uneven class distribution or when both false positives and false negatives are important considerations.

The F1 score is useful in situations where achieving a balance between precision and recall is important, as it penalizes models that have a significant imbalance between these two metrics.

$F1score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

#### 2.1.3.3 In Marketing

**Precision:**

In marketing, precision is valuable when the cost or impact associated with false positives (incorrectly identifying a non-lookalike as a lookalike) is high. For example, if targeting a non-lookalike with a marketing campaign has significant costs, you want to minimize false positives.

**Recall:**

In marketing, recall is important when you want to ensure that you are not missing potential opportunities (actual lookalikes). If missing a true lookalike has a high cost or lost opportunity, you want to maximize recall.

# Chapter 3

# Machine Learning

Different machine learning algorithms are suitable for various types of tasks, such as binary classification, multi-class classification, and predicting continuous outcomes. Here are some commonly used algorithms for each task:

## 3.1 ML Algorithms Intro

### 3.1.1 Binary Classification:

1. **Logistic Regression:**
   - Logistic Regression is a simple and widely used algorithm for binary classification tasks. It models the probability that an instance belongs to a particular class.
2. **Support Vector Machines (SVM):**
   - SVM is effective for binary classification. It finds a hyperplane that best separates the data into two classes.
3. **Random Forest:**
   - Random Forest is an ensemble learning algorithm that performs well for both binary and multi-class classification tasks. It builds multiple decision trees and combines their predictions.
4. **Gradient Boosting (e.g., XGBoost, LightGBM):**
   - Gradient Boosting algorithms are powerful for binary classification tasks. They build trees sequentially, with each tree correcting the errors of the previous one.
5. **Neural Networks:**
   - Neural networks, especially architectures like feedforward neural networks, can be used for binary classification tasks. They are particularly effective for complex, non-linear relationships.

### 3.1.2 Multi-Class Classification:

1. **Logistic Regression (One-vs-All):**
   - Logistic Regression can be extended to handle multi-class classification by training multiple binary classifiers (one for each class) in a one-vs-all fashion.
2. **Multinomial Naive Bayes:**
   - Naive Bayes can be extended to handle multiple classes, and the multinomial variant is commonly used for text classification tasks.
3. **Random Forest:**
   - Random Forest can handle multi-class classification naturally. It builds multiple decision trees, and the final prediction is based on voting across all classes.
4. **Gradient Boosting (e.g., XGBoost, LightGBM):**
   - Gradient Boosting algorithms can handle multi-class classification tasks. They build a series of decision trees, each one correcting the errors of the ensemble.
5. **K-Nearest Neighbors (KNN):**
   - KNN can be used for multi-class classification by assigning the class label that is most common among the k nearest neighbors.

### 3.1.3 Continuous Outcome (Regression):

1. **Linear Regression:**
   - Linear Regression is a basic and widely used algorithm for predicting continuous outcomes. It models the relationship between the features and the target variable as a linear equation.
2. **Decision Trees for Regression:**
   - Decision trees can be used for regression tasks by predicting the average value of the target variable in each leaf node.
3. **Random Forest for Regression:**
   - Random Forest can be applied to regression tasks by aggregating the predictions of multiple decision trees.
4. **Gradient Boosting for Regression (e.g., XGBoost, LightGBM):**
   - Gradient Boosting algorithms can be used for regression tasks. They build decision trees sequentially, each one correcting the errors of the ensemble.
5. **Support Vector Machines (SVR):**
   - Support Vector Machines can be used for regression tasks by finding a hyperplane that best fits the data.

These are just a few examples, and the choice of algorithm depends on factors such as the size and nature of the dataset, the relationship between features and target variables, and computational considerations. It's often a good practice to experiment with multiple algorithms and choose the one that performs best on a specific task.

Several machine learning algorithms are popular and widely used in various applications. The popularity of an algorithm often depends on the nature of the problem and the characteristics of the data. Here are some popular machine learning algorithms:

1. **Linear Regression:**
   - Used for predicting a continuous outcome based on one or more predictor features. It's widely used in regression analysis.
2. **Logistic Regression:**
   - Used for binary classification problems. It models the probability that a given instance belongs to a particular class.
3. **Decision Trees:**
   - A tree-like model of decisions, where each node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label.
4. **Random Forest:**
   - An ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification or the mean prediction for regression.
5. **Support Vector Machines (SVM):**
   - A supervised machine learning algorithm that can be used for classification or regression tasks. It finds a hyperplane that best separates the data into classes.
6. **K-Nearest Neighbors (KNN):**
   - A simple, instance-based learning algorithm where an object is classified by its neighbors. It assigns the class label based on the majority class of its k nearest neighbors.
7. **K-Means Clustering:**
   - A clustering algorithm that partitions data into k clusters based on similarity. It's commonly used for unsupervised learning tasks.
8. **Naive Bayes:**
   - A probabilistic algorithm based on Bayes' theorem that is particularly suited for classification tasks. It assumes that the features are conditionally independent given the class.
9. **Gradient Boosting (e.g., XGBoost, LightGBM):**
   - An ensemble learning technique where weak models (typically decision trees) are trained sequentially, and each new model corrects the errors of the previous ones.
10. **Neural Networks (Deep Learning):**
    - Artificial neural networks inspired by the structure and function of the human brain. Deep learning models, such as feedforward neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs), have achieved remarkable success in various tasks.

These algorithms cover a range of tasks, including regression, classification, clustering, and more. The choice of algorithm depends on the specific problem

at hand and the characteristics of the data. Often, a combination of algorithms or ensemble methods is used to achieve better performance.

### 3.1.4 Random Forest vs Decision Trees

Decision Trees and Random Forest are both machine learning algorithms, and Random Forest is an ensemble learning method that builds on Decision Trees. Here are the key differences between Decision Trees and Random Forest:

#### 3.1.4.1 Decision Trees:

1. **Single Model:**
   - A Decision Tree is a single model that recursively splits the dataset based on the most significant feature at each node.
2. **Vulnerability to Overfitting:**
   - Decision Trees are prone to overfitting, especially when the tree is deep and captures noise in the training data.
3. **High Variance:**
   - Due to their tendency to overfit, Decision Trees have high variance, meaning they can be sensitive to small changes in the training data.
4. **Bias-Variance Tradeoff:**
   - Decision Trees are an example of a model with a high bias (when they are too simple) and high variance (when they are too complex). Finding the right level of complexity is crucial.
5. **Interpretability:**
   - Decision Trees are generally more interpretable, and it's easier to understand the decision-making process at each node.

#### 3.1.4.2 Random Forest:

1. **Ensemble Method:**
   - Random Forest is an ensemble method that builds multiple Decision Trees and combines their predictions. Each tree is trained on a random subset of the data and features.
2. **Reduced Overfitting:**
   - By aggregating predictions from multiple trees, Random Forest reduces overfitting compared to a single Decision Tree. It achieves a better balance between bias and variance.
3. **Improved Generalization:**
   - Random Forest improves generalization performance by creating diverse trees that capture different aspects of the data. The final prediction is an average or a voting mechanism.
4. **Robustness:**
   - Random Forest is more robust to outliers and noisy data compared to a single Decision Tree because the ensemble nature helps filter out noise.

5. **Automatic Feature Selection:**
    - Random Forest provides a form of automatic feature selection by considering a random subset of features at each split in each tree.
6. **Higher Computational Cost:**
    - Building multiple trees and combining their predictions increases the computational cost compared to a single Decision Tree.

In summary, while Decision Trees are simple and interpretable, they are prone to overfitting. Random Forest addresses this limitation by constructing an ensemble of trees, leading to better generalization and robustness at the cost of increased computational complexity. Random Forest is a powerful and widely used algorithm, especially for tasks where high accuracy and robustness are important.

### 3.1.5 Random Forest vs Gradient Boosting

Random Forest and Gradient Boosting are both ensemble learning techniques, but they have some key differences:

#### 3.1.5.1 Random Forest:

1. **Ensemble Type:**
    - Random Forest is an ensemble of decision trees. It builds multiple decision trees independently and combines their predictions through averaging (for regression) or voting (for classification).
2. **Parallel Training:**
    - Trees in a Random Forest can be trained independently and in parallel, making it computationally efficient. This is because each tree is constructed based on a random subset of the data.
3. **Feature Subset at Each Split:**
    - For each split in a tree, a random subset of features is considered. This introduces an element of randomness, reducing the risk of overfitting and making the model more robust.
4. **Voting Mechanism:**
    - In classification tasks, the final prediction is determined by a majority vote from all the individual trees. In regression tasks, the final prediction is the average of the predictions from all trees.
5. **Less Prone to Overfitting:**
    - Random Forest is less prone to overfitting compared to individual decision trees, making it a more robust model.

#### 3.1.5.2 Gradient Boosting:

1. **Ensemble Type:**
    - Gradient Boosting is also an ensemble of decision trees, but unlike Random Forest, it builds trees sequentially, with each tree correcting the errors of the previous one.

2. **Sequential Training:**
   - Trees are trained sequentially, and each subsequent tree focuses on minimizing the errors made by the combined ensemble of the previous trees.
3. **Emphasis on Misclassifications:**
   - Gradient Boosting places more emphasis on correcting the mistakes of the ensemble. Each tree is fitted to the residuals (errors) of the combined model.
4. **Weighted Voting:**
   - In each step, the predictions of all trees are combined with weights, where the weights are determined by the model's performance on the training data.
5. **Potential for Overfitting:**
   - Gradient Boosting has the potential to overfit the training data, especially if the model is too complex or if the learning rate is too high.
6. **More Sensitive to Hyperparameters:**
   - The performance of Gradient Boosting models is more sensitive to hyperparameter tuning compared to Random Forest.

### 3.1.6 Overall Considerations:

- **Parallelization:**
  - Random Forest can be easily parallelized, making it suitable for distributed computing environments.
  - Gradient Boosting, being a sequential process, is not as easily parallelized.
- **Hyperparameter Tuning:**
  - Gradient Boosting typically requires more careful hyperparameter tuning than Random Forest.
- **Performance:**
  - Both models are powerful and widely used, and their performance can vary depending on the characteristics of the dataset.

In summary, while both Random Forest and Gradient Boosting are ensemble methods based on decision trees, they differ in their construction, training process, and emphasis on correcting errors. The choice between them depends on the specific characteristics of the data and the goals of the modeling task.

## 3.2 ML Libraries in Python

Several libraries are widely used for machine learning in addition to scikit-learn. Here are some popular ones:

1. **TensorFlow:**
   - Developed by Google Brain, TensorFlow is an open-source machine

learning library widely used for deep learning applications. It provides a comprehensive set of tools and community support.

2. **PyTorch:**
   - PyTorch, developed by Facebook's AI Research lab (FAIR), is another popular deep learning library. It is known for its dynamic computational graph, making it more flexible for research and experimentation.
3. **Keras:**
   - While Keras can be used as a high-level neural networks API with TensorFlow, it is now also integrated with TensorFlow as its official high-level API. It provides a simple and user-friendly interface for building neural networks.
4. **XGBoost:**
   - XGBoost is an efficient and scalable implementation of gradient boosting. It is widely used for structured/tabular data and is known for its high performance in Kaggle competitions.
5. **LightGBM:**
   - LightGBM is a gradient boosting framework developed by Microsoft. It is designed for distributed and efficient training of large-scale datasets and is particularly useful for categorical features.
6. **CatBoost:**
   - CatBoost is a gradient boosting library that is designed to handle categorical features efficiently. It is developed by Yandex and is known for its ease of use.
7. **Pandas:**
   - While Pandas is not specifically a machine learning library, it is an essential library for data manipulation and analysis. It is often used in the preprocessing phase of machine learning workflows.
8. **NumPy and SciPy:**
   - These libraries are fundamental for scientific computing in Python. NumPy provides support for large, multi-dimensional arrays and matrices, while SciPy builds on NumPy and provides additional functionality for optimization, signal processing, and more.
9. **NLTK and SpaCy:**
   - Natural Language Toolkit (NLTK) and SpaCy are libraries used for natural language processing (NLP). They provide tools for tasks such as tokenization, part-of-speech tagging, and named entity recognition.
10. **Statsmodels:**
    - Statsmodels is a library for estimating and testing statistical models. It is commonly used for statistical analysis and hypothesis testing.

These libraries cover a broad range of machine learning tasks, from traditional machine learning algorithms to deep learning and specialized tools for tasks like natural language processing. The choice of library often depends on the specific requirements of your machine learning project.

### 3.2.1 Big data solutions

When dealing with big data in machine learning, specialized libraries and frameworks that can handle distributed computing and parallel processing become essential. Here are some popular libraries and frameworks for big data machine learning:

1. **Apache Spark MLlib:**
   - Spark MLlib is part of the Apache Spark ecosystem and provides scalable machine learning libraries for Spark. It includes algorithms for classification, regression, clustering, collaborative filtering, and more. Spark's distributed computing capabilities make it well-suited for big data processing.
2. **Dask-ML:**
   - Dask is a parallel computing library in Python that integrates with popular libraries like NumPy, Pandas, and Scikit-Learn. Dask-ML extends Scikit-Learn to support larger-than-memory computations using parallel processing.
3. **H2O.ai:**
   - H2O.ai offers an open-source machine learning platform that includes H2O-3, a distributed machine learning library. H2O-3 supports a variety of machine learning algorithms and is designed to scale horizontally.
4. **MLlib in Apache Flink:**
   - Apache Flink is a stream processing framework, and MLlib is its machine learning library. It allows you to build machine learning pipelines in a streaming environment, making it suitable for real-time analytics on big data.
5. **PySpark (Python API for Apache Spark):**
   - PySpark is the Python API for Apache Spark. It enables Python developers to use Spark for distributed data processing and machine learning tasks. PySpark's MLlib is the machine learning library used within the PySpark ecosystem.
6. **Scikit-Spark (formerly known as BigML):**
   - Scikit-Spark is an extension of Scikit-Learn that allows you to distribute machine learning computations across a cluster. It's built on top of Apache Spark and is designed to handle large datasets.
7. **TensorFlow Extended (TFX):**
   - TFX is an end-to-end platform for deploying production-ready machine learning models at scale. It is built by Google and includes components for data validation, transformation, training, and serving.
8. **Apache Mahout:**
   - Apache Mahout is an open-source project that provides scalable machine learning algorithms. It is designed to work with distributed data processing frameworks like Apache Hadoop.

9. **KNIME Analytics Platform:**
    - KNIME is an open-source platform that allows data scientists to visually design, execute, and reuse machine learning workflows. It supports big data processing through integration with Apache Spark and Hadoop.
10. **Cerebro:**
    - Cerebro is a Python library for distributed machine learning on Apache Spark. It is designed to provide an interface similar to Scikit-Learn for distributed computing.

When working with big data, the choice of library or framework depends on the specific requirements of your project, the characteristics of your data, and the infrastructure you have available. Apache Spark is a particularly popular choice due to its widespread adoption in the big data community.

### 3.2.2 Databricks

Databricks is a cloud-based platform built on top of Apache Spark, and it provides a collaborative environment for big data analytics and machine learning. In Databricks, you have access to various machine learning libraries that integrate seamlessly with Apache Spark. Here are some key machine learning libraries commonly used in Databricks:

1. **MLlib (Spark MLlib):**
    - Apache Spark MLlib is the native machine learning library for Spark. It provides a scalable set of machine learning algorithms and tools, making it a fundamental choice for machine learning tasks in Databricks.
2. **Scikit-learn:**
    - Scikit-learn is a popular machine learning library in Python. While it's not native to Spark, you can use it in Databricks notebooks to perform machine learning tasks on smaller datasets that fit into memory.
3. **XGBoost and LightGBM:**
    - XGBoost and LightGBM are gradient boosting libraries that are widely used for machine learning tasks. They can be integrated with Databricks for boosting algorithms on large-scale datasets.
4. **TensorFlow and PyTorch:**
    - TensorFlow and PyTorch are popular deep learning frameworks. Databricks provides support for these frameworks, allowing you to build and train deep learning models using distributed computing capabilities.
5. **Horovod:**
    - Horovod is a distributed deep learning training framework that works with TensorFlow, PyTorch, and Apache MXNet. It allows you to scale deep learning training across multiple nodes in a Databricks cluster.

6. **Koalas:**
    - Koalas is a Pandas API for Apache Spark, making it easier for data scientists familiar with Pandas to work with large-scale datasets using the Spark infrastructure. It's not a machine learning library itself but can be useful for data preprocessing and exploration.
7. **Delta Lake:**
    - While not a machine learning library, Delta Lake is a storage layer that brings ACID transactions to Apache Spark and big data workloads. It can be used in conjunction with machine learning workflows to manage and version large datasets.
8. **MLflow:**
    - MLflow is an open-source platform for managing the end-to-end machine learning lifecycle. It provides tools for tracking experiments, packaging code into reproducible runs, and sharing and deploying models. MLflow can be easily integrated into Databricks.

When working with Databricks, it's common to leverage MLlib for distributed machine learning tasks and use external libraries like Scikit-learn, TensorFlow, and PyTorch for specific algorithms or deep learning workloads. Additionally, Databricks integrates with MLflow to streamline the machine learning workflow.

## 3.3 Naive Bayes

Naive Bayes models are a group of extremely fast and simple classification algorithms that are often suitable for very high-dimensional datasets. Because they are so fast and have so few tunable parameters, they end up being very useful as a quick-and-dirty baseline for a classification problem.

### 3.3.1 Bayesian Classification

These rely on Bayes's theorem, which is an equation describing the relationship of conditional probabilities of statistical quantities. In Bayesian classification, we're interested in finding the probability of a label given some observed features

#### Gaussian Naive Bayes

Perhaps the easiest naive Bayes classifier to understand is Gaussian naive Bayes. In this classifier, the assumption is that data from each label is drawn from a simple Gaussian distribution. Imagine that you have the following data:

#### When to Use Naive Bayes

Because naive Bayesian classifiers make such stringent assumptions about data, they will generally not perform as well as a more complicated model. That said, they have several advantages:

- They are extremely fast for both training and prediction • They provide

straightforward probabilistic prediction • They are often very easily interpretable • They have very few (if any) tunable parameters

These advantages mean a naive Bayesian classifier is often a good choice as an initial baseline classification. If it performs suitably, then congratulations: you have a very fast, very interpretable classifier for your problem. If it does not perform well, then you can begin exploring more sophisticated models, with some baseline knowledge of how well they should perform.

Naive Bayes classifiers tend to perform especially well in one of the following situations: • When the naive assumptions actually match the data (very rare in practice) • For very well-separated categories, when model complexity is less important • For very high-dimensional data, when model complexity is less important