Murat Durmus

MACHINE LEARNING
Pitfalls

A Brief Guide on How to Avoid Common Pitfalls

(with code examples)

Murat Durmus

Machine Learning Pitfalls

-

A Brief Guide on How to Avoid Common Pitfalls

(With Code Samples)

ISBN-13: 9798386789114

Imprint: Independently published on Amazon (13. March 2023)

Cover design:

Murat Durmus (Mandala: OWL Mandala Mug by Dstudio - canva.com)

About the Author

Murat Durmus is CEO and founder of AISOMA (a Frankfurt am Main (Germany) based company specializing in AI-based technology development and consulting) and Author of the books "Mindful AI - Reflections on Artificial Intelligence" & “A Primer to the 42 Most commonly used Machine Learning Algorithms (With Code Samples) "

You can get in touch with the author via:

- LinkedIn: https://www.linkedin.com/in/ceosaisoma/

- E-Mail: murat.durmus@aisoma.de

Note:

The code examples and their description in this book were written with the support of ChatGPT (OpenAI).

Machine learning is a powerful tool transforming how we approach complex problems, but it is not a panacea.
We can only avoid falling into the traps of bias, overfitting, and unintended consequences by carefully testing and validating the data and algorithms we use.

~ Murat Durmus

Introduction

Machine learning has become ubiquitous in many fields, from healthcare and finance to social media and entertainment. However, as machine learning use grows, so does the importance of understanding its pitfalls. Machine learning models are flexible and can produce unexpected results or even cause harm if not used carefully.

This book is aimed at helping machine learning practitioners, researchers, and enthusiasts avoid the common pitfalls of machine learning. It covers various topics, from data collection and preparation to model selection and evaluation, feature selection and engineering, bias and fairness, interpretability and explainability, and ethical considerations. This book will teach you how to detect and avoid the most common pitfalls in machine learning and build reliable and trustworthy models.

The book is written in a concise and accessible manner and assumes no prior knowledge of machine learning beyond introductory concepts. Each chapter begins with an introduction to the topic, followed by a discussion of common pitfalls, and ends with practical strategies for avoiding those pitfalls. Throughout the book, there is a constant attempt to illustrate the importance of avoiding machine learning pitfalls.

It is not meant to be a comprehensive guide to machine learning or to replace formal education or training in the field. Instead, it is intended to be a practical and accessible guide for anyone interested in using machine learning, regardless of their background or level of expertise.

This book will be a helpful resource for anyone interested in avoiding the pitfalls of machine learning and building trustworthy models. Whether you are a seasoned machine learning practitioner or a newcomer to the field, the lessons in this book will be valuable to you.

Note:

Some techniques and properties, such as confounding variables, use regularization techniques, or Limited / or incomplete data, appear in several areas because they are fundamental clues to pitfalls in machine learning. However, these are always in the context of the respective chapter.

Overview

Chapter 1: *Introduction* The first chapter provides an overview of the book and introduces the concept of machine learning pitfalls.

Chapter 2: *Data Collection and Preparation* This chapter cover common pitfalls related to data collection and preparation, such as data bias, data quality issues, and feature selection.

Chapter 3: *Model Selection and Evaluation* This chapter discusses the importance of choosing the suitable model for a given problem and evaluating the model's performance. In addition, it covers pitfalls related to overfitting, model complexity, and hyperparameter tuning.

Chapter 4: *Feature Selection and Engineering* This chapter focuses on pitfalls related to feature selection and engineering, including feature selection bias, feature engineering mistakes, and dealing with missing data.

Chapter 5: *Bias and Fairness* This chapter covers biases in machine learning models, including algorithmic bias, dataset bias, and bias in model evaluation. It also discusses strategies for detecting and mitigating these biases.

Chapter 6: *Interpretability and Explainability* This chapter discusses the importance of interpretability and explainability in machine learning and covers pitfalls related to model interpretability and explainability.

Why machine learning is prone to pitfalls

Machine learning is prone to pitfalls due to various factors, including the complexity of the models, the quality of the data used to train them, and the biases inherent in the data or the model itself. One of the main reasons machine learning is prone to pitfalls is that it involves building complex models that can be difficult to interpret and understand. In addition, machine learning models are often black boxes, meaning it is not always clear how the model arrives at its predictions. This can make it difficult to detect and diagnose errors or biases in the model.

Another factor that contributes to the pitfalls of machine learning is the quality of the data used to train the model. Machine learning models are only as good as the data they are trained on. If the data is complete, balanced, and of good quality, the resulting model may be reliable and produce accurate predictions. Furthermore, biases inherent in the data or the model itself can affect machine learning models. For example, suppose the data used to train the model is biased in some way, such as by underrepresenting certain groups or perspectives. In that case, the resulting model may produce biased or unfair predictions.

Finally, machine learning models can be vulnerable to overfitting, which occurs when a model is too closely fitted to the training data and does not generalize well to new data. Overfitting can lead to inaccurate predictions and is challenging to detect and correct.

The importance of avoiding pitfalls

Avoiding pitfalls in machine learning is essential for several reasons:

1. **Accuracy**: Pitfalls in machine learning can lead to inaccurate predictions and unreliable models. By avoiding these pitfalls, we can improve the accuracy and reliability of our models.
2. **Transparency**: Machine learning models can often be challenging to interpret and understand. Avoiding pitfalls can help to make models more transparent, allowing us to know how they arrive at their predictions.
3. **Fairness**: Machine learning models can be biased, leading to unfair or discriminatory outcomes. Avoiding pitfalls can help to ensure that models are fair and equitable for all groups.
4. **Trust**: If machine learning models are prone to errors or biases, they can erode trust in the technology. By avoiding pitfalls, we can build more trustworthy models that can be used with confidence.
5. **Ethical considerations**: Machine learning has ethical implications, and avoiding pitfalls can help ensure that our technology use is ethical and responsible.

The importance of high-quality data

High-quality data is essential for building accurate and reliable machine-learning models. Only accurate or complete data can lead to correct predictions, unreliable models, and poor decision-making. Here are some reasons why high-quality data is essential:

1. **Improved Accuracy**: High-quality data can improve the accuracy of machine learning models. Models trained on high-quality data are likelier to make accurate predictions and perform better on test data.
2. **Reliable Insights**: High-quality data provide reliable insights into the problem being modeled. This can help identify patterns, trends, and relationships that inform decision-making.
3. **Better Decision-Making**: High-quality data can lead to better decision-making by providing reliable information for decision-makers. This can help to reduce errors, save time, and improve outcomes.
4. **Reduced Bias**: High-quality data can help to reduce bias in machine learning models. By ensuring that the data is representative and free from discrimination, models can be built that are fair and equitable for all groups.
5. **Cost Savings**: Using high-quality data can save costs by reducing errors and improving decision-making. This can result in lower costs, increased efficiency, and improved outcomes.

Strategies for overcoming data-related pitfalls

Overcoming data-related pitfalls is crucial for building accurate and reliable machine-learning models. Here are some strategies for overcoming data-related pitfalls:

1. **Data Cleaning**: Data cleaning involves identifying and fixing errors and inconsistencies in the data. This can include removing duplicate entries, filling in missing values, and correcting errors in data entry. Data cleaning is crucial in ensuring that the data used to train the model is of high quality.
2. **Data Augmentation**: Data augmentation involves creating new data from existing data by applying transformations such as rotation, flipping, or scaling. This can increase the amount of data available for training the model and improve model performance.
3. **Feature Engineering**: It transforms input data into more valuable features for the modeled problem. This can include creating new features, selecting the most relevant features, and scaling or normalizing features. Feature engineering is an essential step in improving model performance.
4. **Data Balancing**: Data balancing involves addressing data bias by ensuring that the data used to train the model is representative of the population being modeled. This can include oversampling underrepresented groups or undersampling

The importance of choosing the right model

Choosing a suitable model is critical to building effective machine learning systems. The model we choose determines the complexity and expressiveness of the function the machine learning algorithm tries to learn. Choosing the right model can make the difference between a model that performs well and one that performs poorly.

There are several factors to consider when choosing a model:

1. **Complexity**: The complexity of the model should be appropriate for the problem being solved. If the model is more complex, it may fit the data better and perform better. If the model is simple enough, it may overfit the data and perform poorly on new data.
2. **Data Size**: The amount of data available can impact the choice of model. If a large amount of data is available, a more complex model can be used. If a small amount of data is available, a simpler model may be more appropriate.
3. **Interpretability**: The interpretability of the model is essential in some applications. For example, it may be necessary to understand how the model makes predictions in the medical field. In this case, simpler and more interpretable models may be preferred.
4. **Performance Metrics**: The performance metrics used to evaluate the model can also impact the choice of the model. For example, if accuracy is the most

How to detect and avoid overfitting and underfitting

Detecting and avoiding overfitting and underfitting is integral to building machine learning models. Here are some approaches you can use:

1. **Use cross-validation**: Cross-validation is a technique that can help you detect overfitting. By dividing your dataset into multiple subsets and training the model on each subset, you can better understand how well the model generalizes to new data. If the model performs well on the training data but poorly on the test data, this can be a sign of overfitting.
2. **Regularization**: It's a technique that can help prevent overfitting by adding a penalty term to the loss function. This penalty term can encourage the model to have more exact weights, which can help prevent it from overfitting.
3. **Early stopping**: Early stopping is a technique that can help prevent overfitting by stopping the training process when the performance on the validation set stops improving.
4. **Data augmentation**: Data augmentation is a technique that can help prevent underfitting by increasing the size and diversity of the training data. This can help the model recognize more patterns and generalize to new data.

Strategies for avoiding feature-related pitfalls

Some strategies for avoiding feature-related pitfalls:

1. **Understand the problem domain**: To select and engineer the most relevant features, it is essential to have a deep understanding of the problem domain. This involves understanding the underlying factors influencing the outcome and identifying the most relevant variables that capture these factors.
2. **Use domain expertise**: Domain expertise can provide valuable insights into the problem domain and help identify the most relevant features. Experts can help identify the most important variables, suggest appropriate transformations, and provide guidance on which features to use.
3. **Analyze the data**: Analyzing the data can provide insights into the relationship between the features and the target variable. This can involve exploratory data analysis, correlation analysis, and feature importance analysis to identify the most relevant features.
4. **Use regularization techniques**: Regularization techniques such as L1 and L2 regularization can help prevent overfitting and reduce the number of irrelevant features. These techniques penalize complex models, leading to simpler models with fewer features.
5. **Use cross-validation**: Cross-validation can help evaluate the model's performance on independent data and

Hands-on Code Examples

Using LIME to explain the decision of a classification model

Explainable AI (XAI) is an approach to building machine learning models that are transparent and explainable, so that humans can understand the decision-making process of the model. One popular method for XAI is called Local Interpretable Model-Agnostic Explanations (LIME), which explains the model's decision by approximating it with a simpler model that can be easily understood.

```
# import necessary libraries
from sklearn.datasets import load_iris
from sklearn.ensemble import
RandomForestClassifier
import lime
import lime.lime_tabular

# load the iris dataset
data = load_iris()
X = data['data']
y = data['target']
feature_names = data['feature_names']

# train a random forest classifier on the iris
dataset
rfc = RandomForestClassifier(n_estimators=100,
random_state=42)
rfc.fit(X, y)

# create a LIME explainer
explainer =
lime.lime_tabular.LimeTabularExplainer(X,
feature_names=feature_names,
class_names=data['target_names'])
```

```
# choose a data point to explain
idx = 0
x = X[idx]
true_label = y[idx]
print('True label:', true_label)

# explain the model's decision for the chosen
data point
exp = explainer.explain_instance(x,
rfc.predict_proba)

# display the explanation
exp.show_in_notebook(show_table=True,
show_all=False)
```

In this example, we first load the iris dataset and train a random forest classifier on it. We then create a LIME explainer and choose a data point to explain. The **explain_instance** method of the explainer is used to generate an explanation for the model's decision on the chosen data point. Finally, the explanation is displayed using the **show_in_notebook** method.

The output of this code will be an explanation of the model's decision for the chosen data point, showing which features were most important in making the decision. This type of XAI can help humans to better understand the model's decision-making process and identify any potential biases or errors in the model.

Fairness checking

Fairness checking is essential to building fair and ethical machine learning models. One way to perform fairness checking is to use a technique called group fairness, which checks if the model is fair across different groups defined by a protected attribute. Here's an example of how to perform group fairness checking in Python with the Fairlearn library:

```
import numpy as np
import pandas as pd
from sklearn.datasets import
load_breast_cancer
from sklearn.model_selection import
train_test_split
from sklearn.linear_model import
LogisticRegression
from fairlearn.metrics import (
    demographic_parity_difference,
    demographic_parity_ratio,
    equalized_odds_difference,
    equalized_odds_ratio,
)
from fairlearn.postprocessing import
ThresholdOptimizer
from fairlearn.reductions import GridSearch,
DemographicParity

# Load the breast cancer dataset
data = load_breast_cancer()
X = data.data
y = data.target

# Convert the data to a pandas DataFrame
df = pd.DataFrame(X,
columns=data.feature_names)
df['target'] = y
```

```
# Split the data into training and test sets
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=10)

# Train a logistic regression model on the
training data
model = LogisticRegression(random_state=10)
model.fit(X_train, y_train)

# Calculate the demographic parity difference
on the test data
dp_diff = demographic_parity_difference(
    y_test, model.predict(X_test),
sensitive_features=X_test[:, 0] >
np.median(X_test[:, 0])
)
print("Demographic parity difference:",
dp_diff)

# Apply post-processing to mitigate bias
threshold_optimizer =
ThresholdOptimizer(estimator=model,
constraints=DemographicParity(),
objective="accuracy")
threshold_optimizer.fit(X_train, y_train,
sensitive_features=X_train[:, 0] >
np.median(X_train[:, 0]))
y_pred = threshold_optimizer.predict(X_test)
dp_diff_post =
demographic_parity_difference(y_test, y_pred,
sensitive_features=X_test[:, 0] >
np.median(X_test[:, 0]))
print("Demographic parity difference (post-
processing):", dp_diff_post)
```

Typical Stages of Machine Learning Lifecycle

The AI lifecycle is a framework for building and deploying AI systems. It involves several stages: data collection, preprocessing, model development, model training, model testing, deployment, and monitoring. The AI lifecycle aims to ensure that AI systems are developed and deployed consistently and reliably and continue to operate effectively over time.

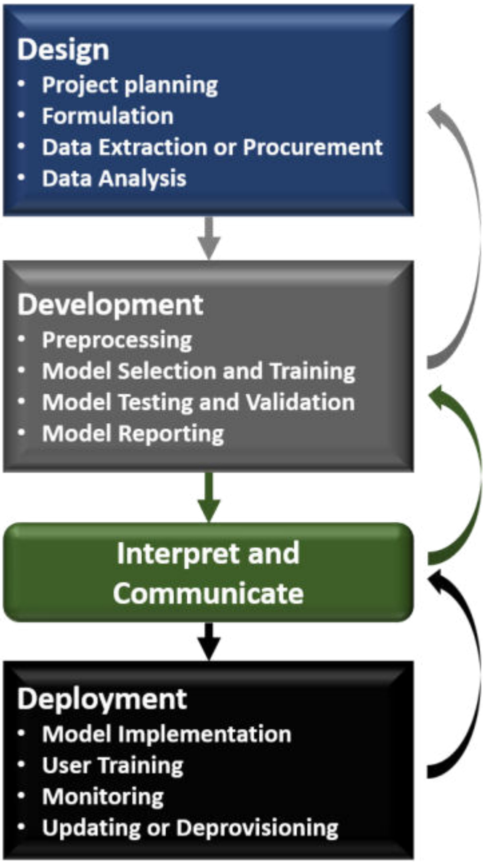


Figure 1: Typical ML Lifecycle

Design

Project Planning

A project team must decide the project's goals at the outset. Tasks in this stage may include stakeholder engagement activities, wider impact assessments, mapping of critical stages within the project, or assessing resources and capabilities within the team or organization. For example, an AI project team is deciding whether or not to use an AI application within an agricultural setting to predict which fields will likely be arable over the next five years and the possible crop yield. This planning allows the project team to reflect on the ethical, socio-economic legal, and technical issues before investing resources into developing the system.

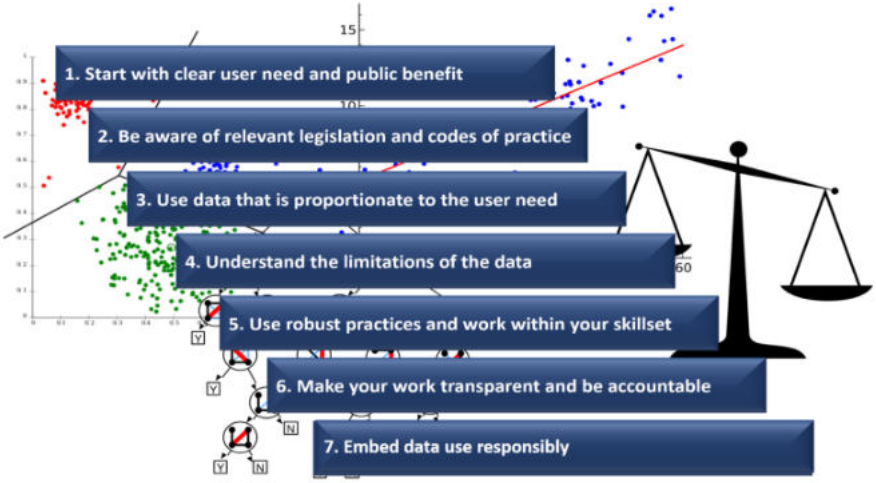
Formulation

A project team needs to determine what problem their model will address and decide what input data is required and for what purpose. The team should consider the ethical and legal implications of the uses of data and provide a thorough account of the intended and unintended consequences of use. For instance, the team has determined the project's overarching theme will involve crop yields. This more precise formulation helps to identify a specific question that can be approached through data and ensure that the result will accord with ethical and legal considerations, such as biodiversity or land use.

Data Extraction or Procurement

This stage involves the processes by which data is gathered for the problem at hand. Data extraction may involve web scraping processes or data recording through surveys or similar methodologies, whereas procurement may involve legal agreements to obtain already existing datasets. In our running example, the team has decided their problem will involve

Data Ethics: A Checklist with 7 Points to Consider



1. Start with clear user need and public benefit
2. Be aware of relevant legislation and codes of practice
3. Use data that is proportionate to the user need
4. Understand the limitations of the data
5. Use robust practices and work within your skillset
6. Make your work transparent and be accountable
7. Embed data use responsibly

Data Ethics refers to the systematization, defense, and recommendation of concepts of right and wrong behavior with data, especially personal data.

Since the dawn of the Internet, the sheer volume and quality of data have increased dramatically, and it continues to grow exponentially. Big Data describes these large volumes of data that are so vast and complex that traditional data processing application software is no longer sufficient to handle them. Recent innovations in medical research and healthcare, such as high-throughput genome sequencing, high-resolution imaging, electronic health records, and a plethora of Internet-connected healthcare devices, have unleashed a deluge of data that will reach the exabyte range shortly. Data ethics becomes more critical as the amount of data increases because of the magnitude of the impact.

The following are seven points to consider:

Privacy in AI systems

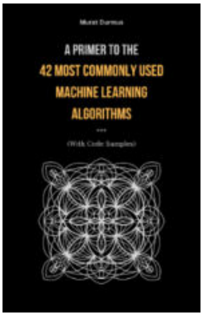
Machine learning models are built on big data and extract statistical patterns to solve a specific task. However, the dataset used for training may be sensitive, either because it contains personal data (medical data, emails, geolocation, etc.) or because the content is restricted (intellectual property, strategic systems, etc.). For personal data, systems should comply with privacy and data protection legislation, and appropriate technical and organizational measures should be taken to implement data protection principles. An example is the EU General Data Protection Regulation (GDPR), which applies in all EU member states. Using anonymization or pseudonymization to this data is a protective measure recommended by the GDPR. However, the feasibility of this measure is highly dependent on the application context and adds to the complexity of the systems deployed, potentially impacting the explainability of the AI system.

When building an AI system based on sensitive data, it is essential to ensure that all actors involved in the machine learning pipeline, from data acquisition to processing, model training, maintenance, and use, are considered trustworthy in handling the data. This section describes the key risks related to data confidentiality and various mechanisms to mitigate these risks and provide some level of privacy protection. These mechanisms are not mutually exclusive and should be combined and complemented with traditional approaches to protecting data management systems.

Also available from the Author

A PRIMER TO THE 42 MOST COMMONLY USED MACHINE LEARNING ALGORITHMS

(WITH CODE SAMPLES)



Whether you're a data scientist, software engineer, or simply interested in learning about machine learning, "A Primer to the 42 Most commonly used Machine Learning Algorithms (With Code Samples)" is an excellent resource for gaining a comprehensive understanding of this exciting field.

Available on Amazon:
https://www.amazon.com/dp/B0BT911HDM

Kindle: **(B0BT8LP2YW)**
Paperback: **(ISBN-13: 979-8375226071)**

MINDFUL AI

Reflections on Artificial Intelligence

Inspirational Thoughts & Quotes on Artificial Intelligence
(Including 13 illustrations, articles & essays for the fundamental understanding of AI)

The field of AI is highly interdisciplinary & evolutionary. The more AI penetrates our life and environment, the more comprehensive the points we have to consider and adapt. Technological developments are far ahead of ethical & philosophical interpretations; this fact is disturbing.

We need to close this gap as soon as possible.

~ (Mindful AI)

MINDFUL AI
Reflections on Artificial Intelligence
(THE AI THOUGHT BOOK Reloaded)

Available on Amazon:
https://www.amazon.com/dp/B0BKMK6HLJ

Kindle: **(ASIN: B0BKLCKM22)**
Paperback: **(ISBN-13: 979-8360396796)–**