Table of contents

Machine Learning and Natural Language Processing

Student Edition

By Kimberly Seefeld, EdD

Published under **DataJoyAI** imprint

Companion resources for this book are available at:

https://datajoyai.com

Edition: First Edition
ISBN: 978-1-969233-32-6

Chapter 1

Introduction to Applied Machine Learning

1.1 Learning Outcomes

1-1: Distinguish between supervised and unsupervised learning approaches and identify appropriate applications

1-2: Design and implement a complete machine learning pipeline from data preprocessing through model evaluation

1-3: Evaluate model interpretability requirements and implement bias detection strategies

1-4: Select and configure AI-assisted model selection tools to optimize performance

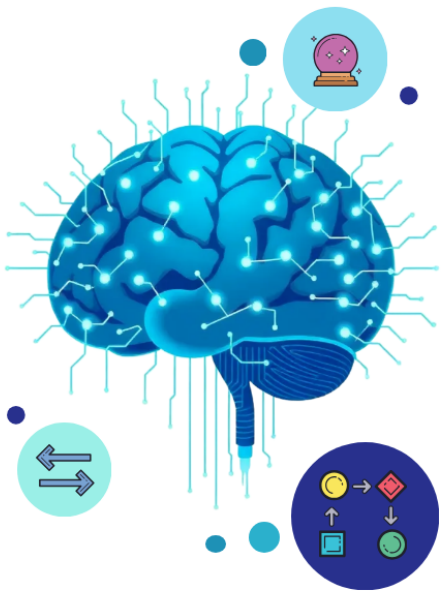
1.2 Introduction

Picture this scenario: You're scrolling through Netflix, and within seconds, the platform presents you with a curated list of shows that seem perfectly tailored to your taste. Meanwhile, your bank's fraud detection system is silently analyzing thousands of transactions per second, flagging suspicious activity before you even notice. In another part of the world, a radiologist is reviewing medical images with the assistance of an AI system that can detect early-stage cancer with remarkable accuracy. What connects these seemingly different experiences? They all represent the power of applied machine learning transforming how we live, work, and make decisions.

But let's step back for a moment. Before you even grabbed your phone this morning, you were already immersed in a world shaped by artificial intelligence. Your alarm clock learned your sleep patterns to wake you at the optimal time. Your coffee maker started brewing because it predicted when you'd wake up. Your car's navigation system analyzed traffic patterns to suggest the best route to campus. By the time you check Instagram and see posts from friends you actually care about at the top of your feed, you've already interacted with machine learning algorithms dozens of times.

This book focuses on understanding how machine learning moves from theoretical concepts to real-world business solutions. However, before we dive into advanced applications, we need to establish a solid foundation in the fundamental approaches that drive these systems. This opening module introduces the core concepts that every data professional must master: what machine learning actually is and how it fits within the broader landscape of artificial intelligence, the distinction between supervised and unsupervised learning, the systematic pipeline that transforms raw data into actionable insights, and the critical considerations of model interpretability and bias that determine whether our solutions succeed or fail in practice.

1.3 Understanding What Machine Learning Is



1.3.1 The Artificial Intelligence Landscape

Before diving into machine learning specifics, we need to understand where it fits in the bigger picture of artificial intelligence. Think of AI as a vast umbrella that covers any technology designed to make machines seem intelligent or human-like. Under this umbrella, you'll find several different approaches to creating smart systems, each with its own strengths and applications.

Artificial Intelligence represents the broad field focused on creating machines that can perform tasks typically requiring human intelligence. This encomtest set. This gives a concrete sense of how the models' predictions compare to the actual spending values. It helps students see the real-world impact of model choice and feature selection.

```
# Show some example predictions
print("\nExample Predictions (first 10 test customers):")
print("-" * 50)
comparison_df = pd.DataFrame({
    'Actual': y_test.iloc[:10].values,
    'Linear_Reg': y_pred_linear[:10],
    'Decision_Tree': y_pred_tree[:10],
    'Random_Forest': y_pred_rf[:10]
})
comparison_df = comparison_df.round(2)
print(comparison_df)
```

Output:

```
Example Predictions (first 10 test customers):
--------------------------------------------------
   Actual  Linear_Reg  Decision_Tree  Random_Forest
0  291.00      252.44         270.64         269.25
1  180.83      224.72         243.12         236.42
2  494.51      477.18         511.35         416.80
3  257.70      275.39         279.85         289.02
4  313.39      278.37         274.52         302.99
5  232.77      206.85         156.01         199.31
6  132.02      170.43         201.50         171.01
7  221.19      153.01         201.50         173.73
8  867.37      728.20         615.51         576.26
9  265.05      207.95         156.01         193.00
```

2.11.2.4 Step 4: Classification - Predicting Purchase Behavior

Now let's solve our second problem: predicting whether a customer will make a purchase this month. This is a classification problem because we're predict-

```
    (shipping_method == 'Overnight').astype(int) * 1.0 +
    np.random.normal(0, 1, n_customers)
)
satisfaction_score = np.clip(satisfaction_score, 1, 10)

customer_data = pd.DataFrame({
    'age': age,
    'order_value': order_value,
    'delivery_days': delivery_days,
    'num_items': num_items,
    'customer_service_calls': customer_service_calls,
    'shipping_method': shipping_method,
    'product_category': product_category,
    'payment_method': payment_method,
    'satisfaction_score': satisfaction_score
})

print(f"Dataset created with {len(customer_data)} customers")
print(customer_data.head(10))
```

2.12.3.1 Lab 2B Questions

1. How many customers are in the dataset?
2. What is the average satisfaction score for all customers?
3. Which shipping method is the most common?
4. Which product category appears most frequently?
5. Build a basic Linear Regression model using the numerical features (age, order_value, delivery_days, num_items, customer_service_calls) to predict satisfaction_score. What is the R^2 score?
6. Which feature has the largest positive effect on satisfaction in your basic model?
7. Which feature has the largest negative effect on satisfaction in your basic model?

Chapter 3

Model Evaluation and Performance Metrics

3.1 Learning Outcomes

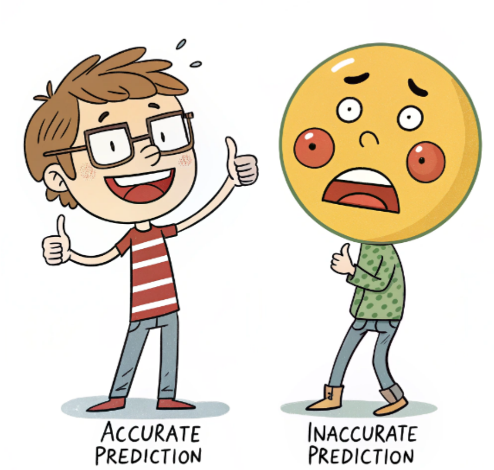
3-1: Understand different ways to measure model performance and choose the right metrics for different business problems

3-2: Explain what over fitting means and how cross-validation helps create more reliable models

3-3: Recognize when models might be biased or unfair and understand basic approaches to evaluation

3-4: Interpret common evaluation tools like confusion matrices and performance charts

3.2 Introduction



Imagine you've just built your first machine learning model to predict which customers will cancel their subscriptions. You're excited—the model seems to work great! It correctly predicts 95% of the cases in your test data. You present your results to your boss, who asks a simple question: "How do you know it's actually good?"

This question might seem obvious, but it reveals one of the most critical aspects of machine learning that beginners often overlook. Building a model is only half the battle—the other half is figuring out whether it actually works well enough to trust with real business decisions. After all, a model that seems accurate in testing but fails in the real world can cost companies millions of dollars and damage customer relationships.

Consider what happened to a major retailer that built a model to predict product demand. In testing, the model achieved 90% accuracy, so they deployed it to manage inventory for the holiday season. But the model had a hidden flaw—it was terrible at predicting demand for new products, which made up 30% of their holiday inventory. The result? Massive overstocking of some items and shortages of others, leading to millions in losses.

This module will teach you how to properly evaluate machine learning models so you can avoid these costly mistakes. We'll explore the different ways to measure model performance, understand why simple accuracy isn't always enough,

Chapter 4

Introduction to Natural Language Processing

4.1 Learning Outcomes

4-1: Understand what Natural Language Processing is and why it's important for modern businesses

4-2: Explain the fundamental challenges of processing human language with computers and the basic approaches to preparing text for analysis

4-3: Describe how text is converted into numerical representations that machine learning algorithms can work with

4-4: Recognize common NLP applications and understand the basic techniques that power them grouping makes linguistic sense.

4.6 Converting Text to Numbers

Machine learning algorithms work with numbers, not words. To apply machine learning techniques to text, we need methods for converting words and documents into numerical representations that capture their meaning and relationships. The challenge isn't just converting text to numbers—it's doing so in a way that preserves meaning.

4.6.1 TF-IDF: Measuring Word Importance

Term Frequency-Inverse Document Frequency (TF-IDF) is one of the most important and widely used techniques for converting text to numbers. It measures how important a word is to a document within a collection of documents.

TF-IDF is based on two simple ideas: words that appear frequently in a document are important to that document, and words that appear in many documents are less distinctive and therefore less important. Think about it this way: if the word "election" appears many times in a news article, it's probably important to that article. But if "election" appears in every political article, it's not very useful for distinguishing one political article from another.

Term Frequency measures how often a word appears in a document, usually normalized by the total number of words in the document. In a 100-word document, if the word "customer" appears 5 times, its TF score is 0.05. This normalization is important because longer documents naturally contain more words,

```
neutral     478
negative    453
Name: count, dtype: int64

Rating distribution:
rating
1    270
2    183
3    466
4    156
5    425
Name: count, dtype: int64
```

4.9.2.2 Step 2: Text Preprocessing - Cleaning and Preparing Text

Before analyzing text data, it is essential to clean and standardize it. Raw text often contains inconsistent capitalization, punctuation, extra whitespace, and words that add little meaning, which can make analysis less accurate or even misleading. Preprocessing transforms this raw text into a consistent format that machine learning algorithms or rule-based systems can interpret reliably.

Text preprocessing typically involves several key steps. First, basic cleaning converts all text to lowercase, removes punctuation, and normalizes whitespace. This ensures that words like "Good" and "good" are treated the same and that extra spaces or symbols do not interfere with analysis. Next is tokenization, which splits text into individual words or tokens. Tokenization allows us to examine or model the text at the word level. Stopword removal is often applied to eliminate extremely common words, like "the" or "and," that do not contribute meaningful information for sentiment or topic analysis. Finally, feature extraction converts the clean text into numerical representations, such as TF-IDF vectors, which can be used by machine learning models.

The code implements these preprocessing steps. The `clean_text` function handles basic cleaning by converting text to lowercase, removing punctuation, and removing extra spaces. The `preprocess_text` function is a slightly more advanced version, using regular expressions to carefully remove punctuation while preserving word spacing. These functions ensure that the text is in a conyou receive hundreds of online reviews every week. Some reviews say "The food was absolutely amazing and the service was perfect!" while others complain "Terrible experience, cold food and rude staff." As a human, you immediately understand that the first review expresses positive feelings while the second expresses negative feelings. Sentiment analysis teaches computers to make these same distinctions automatically.

Sentiment analysis is the process of automatically determining the emotional tone behind text—whether someone is expressing positive, negative, or neutral feelings about a topic. This capability has become essential for businesses trying to understand customer satisfaction, monitor brand reputation, and respond appropriately to customer feedback in our digital world.

The power of sentiment analysis becomes clear when you consider the scale of modern communication. A large company might receive thousands of customer emails, social media mentions, and product reviews every day. Reading through all this feedback manually would take weeks, but sentiment analysis can process it in minutes, automatically flagging the most urgent negative feedback while identifying positive trends worth celebrating.

5.3.1 How Sentiment Analysis Works

At its core, sentiment analysis is about teaching computers to recognize emotions in text. Just like humans can tell when a sentence expresses happiness, anger, or disappointment, sentiment analysis systems learn to associate certain words and phrases with positive, negative, or neutral feelings. Words like "excellent," "amazing," and "love" generally indicate positive sentiment, while "terrible," "awful," and "hate" usually signal negative sentiment. These sys-