

Introduction to Machine Learning and Finance

Prepared by Dr. Sumaya Alkubaisi

FNT633

Machine Learning in Finance

Lecture 1



Chapter

1

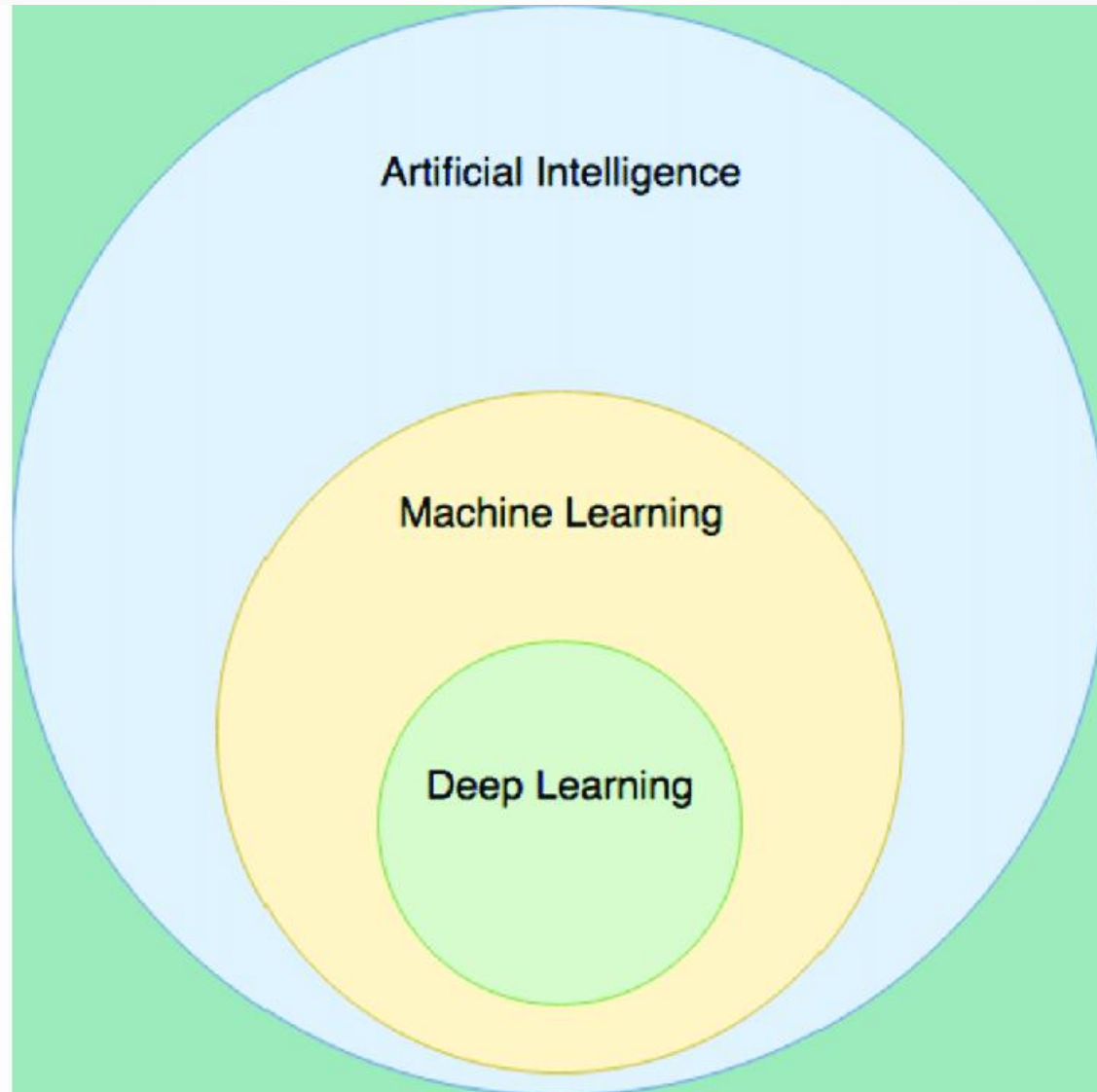
Learning Goals

By the end of this lecture, you will be able to:

- Understand what Machine Learning (ML) is and why it matters in finance.
- Describe the Data Science Life Cycle and its steps.
- Explain the importance of Data Understanding before modeling.
- Identify financial data types (numerical, categorical, time series).
- Recognize common financial data sources (market, credit, transaction).
- Write simple Python code for data loading, cleaning.

What is Machine Learning?

- Machine Learning (ML) is a branch of Artificial Intelligence (AI) where computers learn from data instead of being programmed with fixed rules.
- ML algorithms identify patterns in data and make predictions or decisions.
- In finance, ML helps automate tasks, forecast markets, and manage risks.



- **AI (Artificial Intelligence):** The big umbrella — making machines act smart — able to mimic human thinking, reasoning, and decision-making.
- **ML (Machine Learning):** A subset of AI — machines learn from data to improve without being told every rule.
- **Deep Learning (DL):** A subset of ML — uses artificial neural networks (like the human brain's structure) to handle very complex tasks such as image recognition, speech, and natural language.

Why ML in Finance?

Speed – Automates decisions in milliseconds

ML algorithms can process millions of transactions instantly.

Example: High-frequency trading (HFT) systems use ML to buy and sell stocks in fractions of a second, exploiting tiny price differences before humans can react.

Accuracy – Detects fraud better than manual checks

ML models learn patterns of legitimate vs. fraudulent transactions.

Example: Credit card companies use ML to flag unusual spending behavior, preventing fraudulent purchases automatically.

Why ML in Finance?

Insights – Finds hidden market patterns

ML uncovers trends and correlations that traditional analysis might miss.

Example: Portfolio managers use ML to detect correlations between asset classes, helping to predict market movements and optimize investment strategies.

Efficiency – Reduces cost and human error

Automating repetitive tasks saves time and reduces mistakes.

Example: Loan approval processes use ML to assess risk profiles faster than manual review, reducing operational costs and errors in decision-making.

ML Example in Finance

Predicting Loan Default

ML models analyze historical loan data (income, credit score, debt, etc.) to predict if a borrower might default.

Example: A bank uses a Random Forest model to assign a “default risk score” to each applicant, helping decide loan approval automatically.

Forecasting Stock Prices

ML algorithms detect patterns in historical stock prices, trading volumes, and market indicators to predict future prices.

Example: LSTM (Long Short-Term Memory) networks are used to forecast daily stock prices, helping traders make informed investment decisions.

ML Example in Finance

Detecting Fraudulent Transactions

ML identifies unusual patterns in credit card or transaction data that indicate fraud.

Example: An online payment platform uses anomaly detection to flag transactions that deviate from normal spending behavior, alerting customers in real-time.

Data Science Life Cycle

What is the Data Science Life Cycle?

- The Data Science Life Cycle is a structured, systematic process used to extract knowledge, insights, and actionable solutions from data.
- It ensures that data-driven projects are organized, repeatable, and reliable, reducing errors and improving decision-making.
- This lifecycle guides practitioners from understanding the problem and collecting raw data, through preparing and analyzing it, to deploying models and monitoring their performance in real-world scenarios.

Life Cycle Stages

Main phases: Collect → Prepare → Analyze → Evaluate → Deploy.

- Where Analyze means modeling (training and testing models).
- **Data Collection** – Gather raw relevant data from multiple sources
Example: A bank collects historical loan data, customer demographics, and transaction records.
- **Data Preparation** – Clean, integrate, and transform data for analysis
 - Handling missing values, removing outliers, normalizing data, and encoding categorical variables.Example: Filling missing income values with median values and encoding “Yes/No” columns as 1/0.

Life Cycle Stages

Modeling – Train and test machine learning models

Use statistical methods and ML algorithms to extract insights or make predictions.

Example: Training a logistic regression model to predict loan default, and testing its accuracy on unseen data.

Evaluation – Measure performance and validate results

Check model accuracy, precision, recall, or other relevant metrics.

Example: Evaluating fraud detection model using a confusion matrix to see how many fraudulent transactions were correctly identified.

Deployment – Implement the model in real-world applications

Integrate the solution into production systems for continuous use.

Example: Deploying a credit scoring model to automatically approve or flag loan applications in the bank's system.

Example in Finance

Problem: Predict if a loan applicant will default.

Steps:

Steps:

- **Collect** – Gather relevant data

Example: Applicant's income, credit history, existing debts, age, employment status.

- **Prepare** – Clean and organize data

- Handle missing values, remove errors, standardize formats, and encode categorical data.

Example: Fill missing income values with median income, convert “Yes/No” fields to 1/0.

Example in Finance

- **Model** – Train a machine learning algorithm
 - Use historical loan data to teach the model patterns of default vs. successful repayment.
Example: Train a Random Forest or Logistic Regression model to predict default risk.
- **Evaluate** – Test how well the model works
 - Measure accuracy, precision, recall, and other metrics to ensure predictions are reliable.
Example: Check how many defaulters the model correctly predicts versus false positives.
- **Deploy** – Apply the model in real life
 - Integrate the model into the bank's system to automatically assess new loan applications.
Example: Assign a "default risk score" to applicants to support faster, data-driven approval decisions.

Data Collection in Finance

Sources of Financial Data:

Stock market prices: Historical and real-time data from exchanges.

Customer bank records: Transactions, balances, and account activity.

Credit bureau data: Credit scores, repayment history, loan applications.

Online transactions: E-commerce and digital payments data.

Data Collection in Finance

Practical Note:

- Access to real financial data is often restricted due to privacy and regulations.
- For learning and experimentation, we use public datasets from platforms like **Kaggle and UCI Machine Learning Repository**, for example:
 - Credit Card Fraud Detection
 - Loan Default Prediction
 - Stock Price Data

Data Science Life Cycle in Python

#Python Setup

```
import pandas as pd
```

- pandas: manage financial data
- allows us to read, clean, and manage financial data easily, whether it comes from Excel files, CSVs, or online sources.

Data Science Life Cycle in Python

#Data collection: Load Data in Python

```
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-ML0187EN-SkillsNetwork/labs/module%202/iris.csv")  
df.head()
```

- `pd.read_csv()` - Loads CSV into a DataFrame.
- `head()` - shows first 5 rows.

Data Science Life Cycle in Python

#Clean Data in Python

```
df.drop(['petal_width'], axis=1, inplace=True)  
df.head()
```

- `df.drop()` - Removes unwanted column.

Data Understanding

What is Data Understanding?

The step of exploring and analyzing raw data before building models.

Involves looking at the data to see what it contains, how it is structured, and whether it makes sense.

Helps to detect errors, missing values, outliers, and unusual patterns that could mislead the model.

Ensures that the dataset is relevant, accurate, and meaningful for the problem being solved.

Example: In a loan dataset, checking if any applicant has a negative income value (error) or if many records are missing credit history (incomplete data).

Lecture

1

Key Terms in ML – Dataset, Features, Target

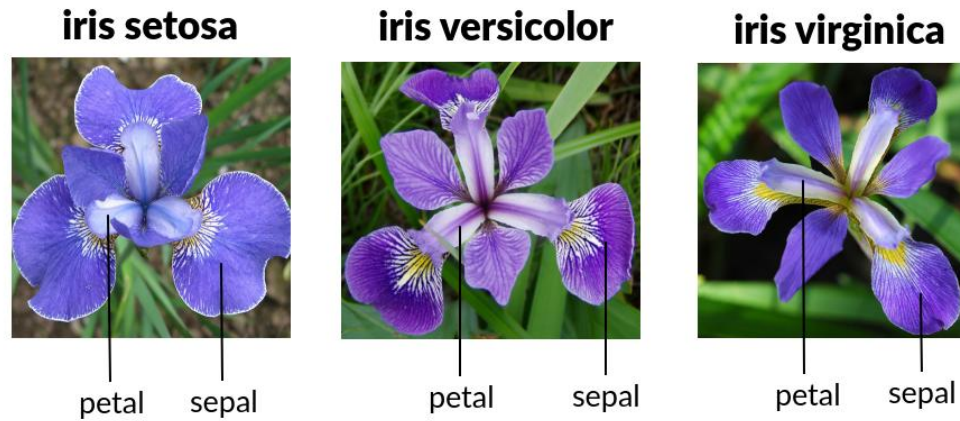
- **Dataset** – A structured collection of data used for analysis or to train a machine learning model.
 - Usually organized in rows and columns, like a table.
 - Example: A loan application dataset where each row represents one applicant, and each column represents their details (income, age, credit score, etc.).
- **Features (Input Variables)** – The information or characteristics used by the model to make predictions.
 - These are the independent variables.
 - Example: Applicant's income, employment status, age, loan amount, and credit score.

Key Terms in ML – Dataset, Features, Target

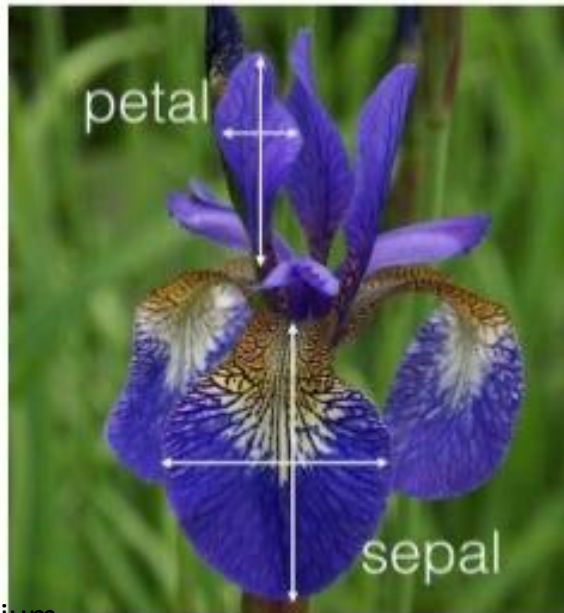
- **Target (Label / Output Variable)** – The outcome we want the model to predict.
-This is the dependent variable.
Example: Whether the loan was approved (“Yes/No”) or whether the applicant defaulted (“Default/No Default”).

In short: Features go in → Target comes out.

Image source: Code signal learn



Supervised learning *classification* problem (using the [Iris flower data set](#))



Training / test data

Features Labels

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

Image source: Medium

Iris Flower Dataset

Iris Dataset Overview

What is the dataset?

A famous flower dataset used in ML, containing measurements of Iris flowers.

Variables in the dataset:

Features (Inputs):

Sepal length, Sepal width

Petal length, Petal width

Target / Label (Output):

Species → Iris setosa, Iris versicolor, or Iris virginica

Relation: Features describe flower measurements, while the target identifies the flower type.

Iris Flower Dataset

How ML Works with Iris Data?

Training step:

The ML model learns the relationship between flower measurements and their species.

Example: Iris setosa usually has small petals, while Iris virginica has larger petals.

Prediction step:

New flower measurements are provided.

The model predicts the species.

Use case: Helps botanists or researchers automatically classify flower species based on measurements.

Record #	Default	Num_Loans	Amount	Duration	Age
1	Yes	1	1295	12	25
2	Yes	1	4308	48	24
3	Yes	1	12579	24	44
4	No	1	1804	12	44
5	No	2	409	12	42

Image source: Alteryx Community

Loan Default Dataset

What is the Dataset?

A collection of loan applications from customers, with information about their loan and whether they repaid (No Default) or failed to repay (Default).

Each **row** = one customer's loan record.

Columns (variables):

Record # (ID of the record)

Default (Yes/No) → did the customer repay?

Num_Loans → number of loans taken

Amount → loan amount requested

Duration → repayment period in months

Age → age of the customer

Loan Default Dataset (Features, Target, Inputs, and Outputs)

Features (Inputs to ML model):

Num_Loans, Amount, Duration, Age

Target / Label (Output we want to predict):

Default (Yes = defaulted, No = repaid)

Type of problem:

Classification → Predict category (Yes or No).

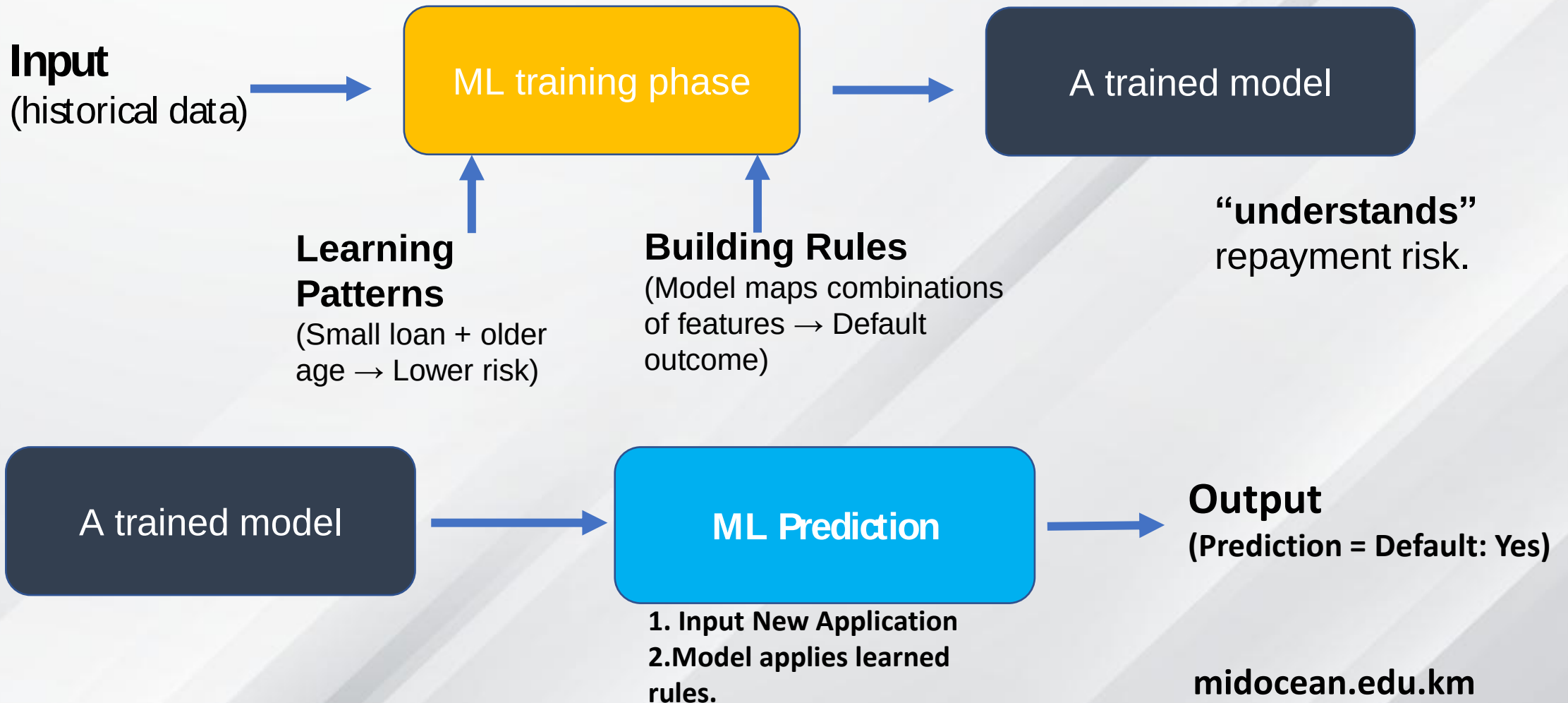
Relations:

Features describe the customer's profile and loan details.

Target tells us the final outcome (Default or No Default).

How ML Uses This Dataset (Training Phase)

1



Summarizing Data

Descriptive Statistics (mean, min, max, quartiles.)

Shows key numbers that summarize the dataset, such as:

Mean (Average) – the central value of the data.

Useful for: Understanding the typical or expected value in the dataset.

Example: Average applicant income = 50,000 helps the bank know the usual income range of clients.

Minimum & Maximum – the lowest and highest values in the data.

Useful for: Detecting extreme values and checking data validity.

Example: If minimum income = -5,000 → data entry error; if maximum income = 1,000,000 → possible outlier.

Summarizing Data

Quartiles (Q1, Median, Q3) – divide data into four equal parts to show distribution.

Useful for: Understanding spread, balance, and skewness of data.

Example: If $Q1 = 30,000$, $Median = 50,000$, $Q3 = 70,000$ → most applicants earn between 30k–70k, with half below 50k and half above.

Example:

In a loan dataset, if the income column shows:

- Mean (Average applicant income) = 50,000 (tells the bank the usual income of clients.)
- Min = –5,000 (clearly an error, negative income, income can't be negative)
- Max = 1,000,000 (unusually high, possible outlier, needs checking)
- $Q1 = 30,000$, $Median = 50,000$, $Q3 = 70,000$ → most applicants earn between 30k and 70k, with half earning less than 50k and half more.

Data Visualization

Charts reveal patterns quickly.

- Easier to understand than looking at raw numbers.
- Help spot trends, clusters, outliers, and distributions at a glance.
- Useful for both exploring data and explaining insights to others.

Examples:

1. Stock Price Trend Over Time (Line Chart)

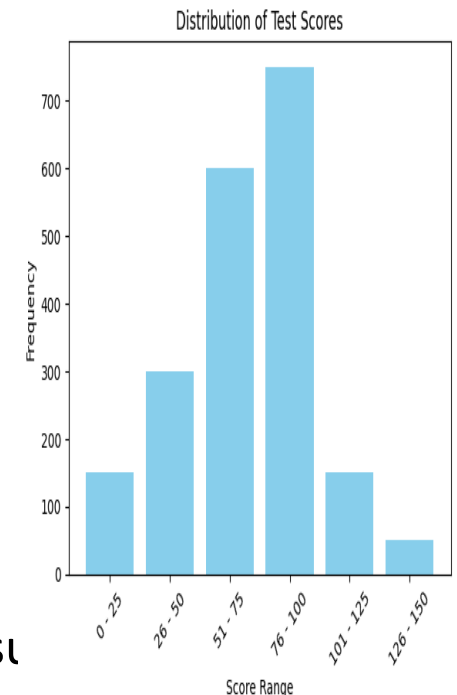
Shows how stock values rise and fall across days, months, or years.

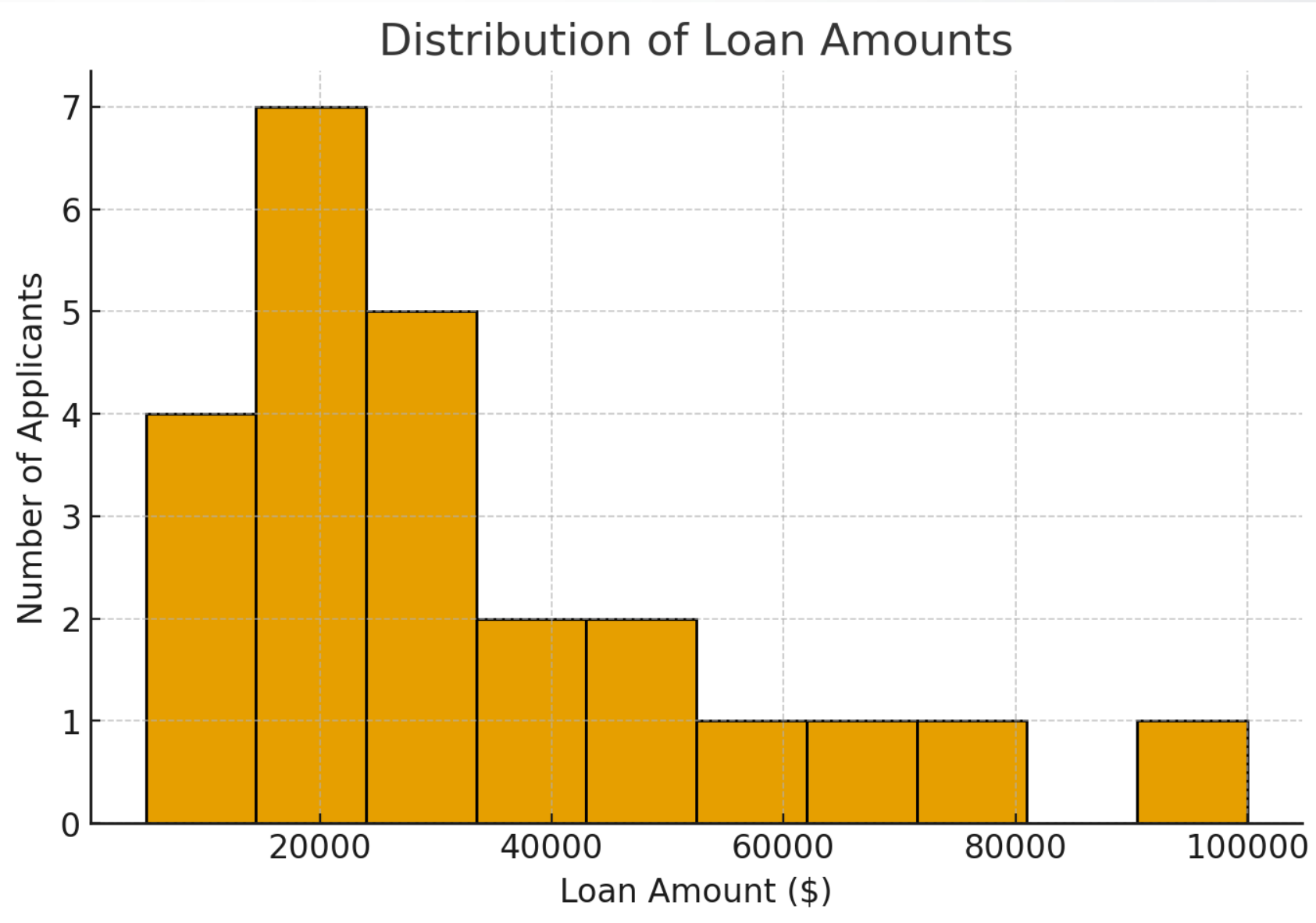
Useful for: Spotting upward/downward trends, sudden drops, or unusual spikes.

2. Distribution of Loan Amounts (Histogram / Bar Chart)

Displays how many applicants fall into different loan size ranges.

Useful for: Checking if most loans are small, medium, or large, and spotting unus





Data Types in Finance

Numerical: Data expressed in numbers, such as prices, loan amounts, or interest rates.

Example: **Numerical:** Salary = 5000 USD.

Categorical: Data divided into groups or categories (classified data), such as loan type (personal, mortgage), credit rating (A, B, C).

Example: **Categorical:** Risk level = “High.”

Time Series: Data recorded in sequence over time, such as stock prices or daily exchange rates.

Example: **Time Series:** Apple stock closing prices over 1 year.

Correlation in Data

Correlation = Relationship between two variables.

Range:

+1 (positive correlation): When one goes up, the other also goes up.

-1 (negative correlation): When one goes up, the other goes down.

0 (no correlation): No clear relationship.

Example in finance:

Income $\uparrow$ $\rightarrow$ Loan approval chance $\uparrow$.

Data Understanding in Python

#Summarizing Data in Python

`df.describe()`

- Shows statistics: mean, min, max, quartiles.
- Useful for spotting outliers or strange values.

Data Understanding in Python

#Check Data Types in Python

`df.dtypes`

Lists each column's type (int, float, object, etc.).

Data Understanding in Python

#Correlation Matrix in Python

`df.corr()`

- Creates correlation values for all features.
- Helps select the most useful features for ML.

Financial Data Types and Sources

Types of Financial Data

- **Market Data** – Stocks, bonds, forex, options.
- **Credit Data** – Loans, repayment history, credit scores.
- **Transaction Data** – Purchases, bank transfers, credit card usage.

Market Data Example

- Stock market data includes:
 - Open, High, Low, Close prices.
 - Daily trading volume.
- Sources of Market Data
 - Bloomberg (professional financial terminals).
 - Yahoo Finance (public stock data).
 - Kaggle datasets (academic/learning datasets).

Transaction Data

- Includes credit/debit card transactions.
- Online payments, e-commerce logs.
- Used for fraud detection.

Example of Fraud Detection

- If a card is used in two countries within one hour → suspicious.
- ML models flag unusual spending behavior.

Lecture

1

Real Industry Use Case

- Visa and Mastercard use ML fraud detection systems.
- Models scan millions of transactions per second.
- Alerts are raised in real time.

Summary: Financial Data Sources

- **Market data** → stocks, forex.
 - **Credit data** → loans, scores.
 - **Transaction data** → purchases.
- ❖ All are essential for ML in finance.

Lecture

1

Recap

In this lecture, we learned:

- **Machine Learning:** Computers learn from data to predict/decide.
- **Finance Applications:** Trading, fraud detection, credit scoring.
- **Data Science Life Cycle:** Collect → Prepare → Model → Evaluate → Deploy.
- **Data Understanding:** Explore data before modeling (types, patterns, statistics).
- **Financial Data:** Market, credit, transaction.
- **Python Practice:** Loaded and cleaned data.

References

- Machine Learning for Finance (Jannes Klaas).
- Advances in Financial Machine Learning (López de Prado).