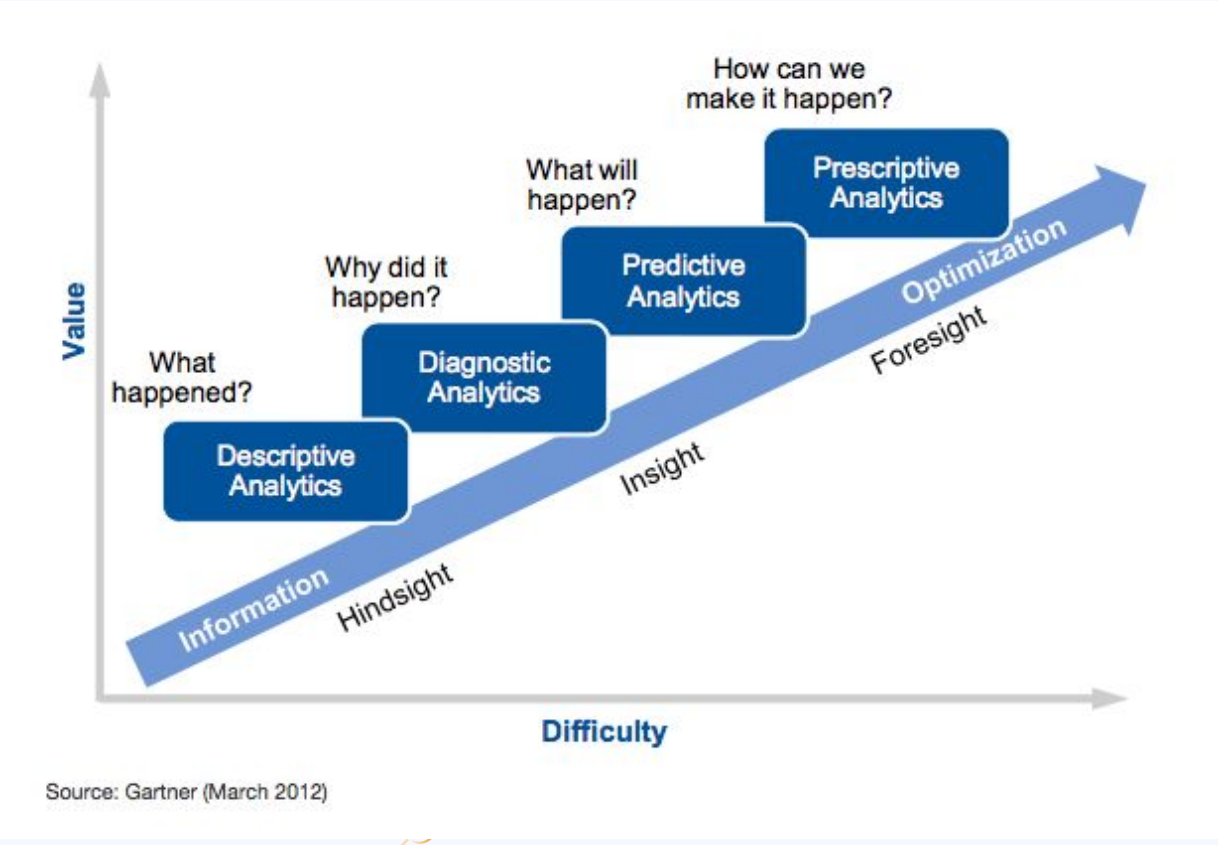
Predictive Analytics

Leonard Andrew Mesiera,MSIT
DLL - Bachelor of Science in Information Technology
AY 2025 - 2026
ITELEC303

The Data Evolution:

From Looking Back to Looking Forward

Stages of Data Analytics



Source: Gartner (March 2012)

Predictive Analytics

In simpler terms, it answers the question: **"What *might* happen?"**

The branch of advanced analytics that uses current and historical data, combined with statistical modeling, data mining techniques, and machine learning, to forecast the probability of future outcomes, events, or behaviors.

Key Components of Predictive Analytics:

1. **Historical Data:** This is the foundation. Predictive models learn from past patterns (e.g., past customer purchases, machine failures, or loan defaults).
2. **Advanced Techniques:** It relies on mathematical models and algorithms (often Machine Learning) to find relationships within the data.
3. **Probabilistic Output:** The output is typically a **score or a probability** (e.g., "This customer has an $85% likelihood of defaulting," or "Demand for this product is predicted to be 5,000 units next month"). It forecasts a *likelihood*, not a certainty.
4. **Purpose:** To enable **proactive decision-making**. By anticipating future trends, organizations can mitigate risks (like fraud or equipment failure) and capitalize on opportunities (like targeted marketing or demand forecasting).

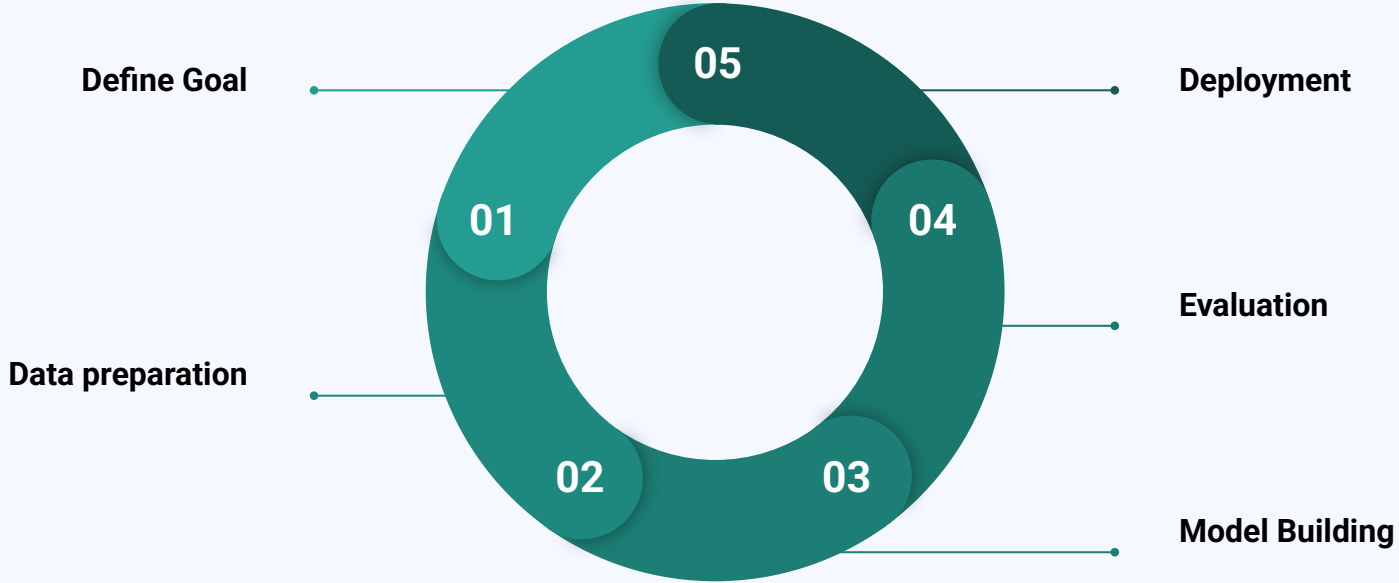
How can someone make prediction ?

Concept	Explanation for Prediction	Analogy
It's not Magic, It's Math.	A prediction is a **calculated probability**, not a guaranteed certainty. We determine the *likelihood* of a future event based on patterns observed in the past.	**The Weather Forecast:** Meteorologists don't *know* it will rain; they calculate the high probability based on current atmospheric conditions, temperature, pressure, and historical storm data.
The Core Question	Given a set of input variables (X_1, X_2, X_3, ...) what is the most probable output (Y) in the future?	**Past Customer Data** → Prediction of Churn Likelihood
Formula (Conceptual)	Future Outcome = Model x (Historical Data + Current Conditions)	

The Core Science:

Statistics, Data Mining, and Modeling

The Predictive Process: A 5-Step Workflow



Variables: The Language of Prediction 📊

Variable	TypeRole in Prediction	What it Answers	Analogy
Independent Variable (Input/Predictor)	**The Cause** or **The Factor** that is measured, controlled, or manipulated. The model uses this to make a forecast.	What do we know?	The **ingredients** in a recipe (Sugar, Flour, Time in Oven).
Dependent Variable (Output/Target)	**The Effect** or **The Outcome** that the model is designed to predict. Its value depends on the independent variable(s).	What do we want to predict?	The **final outcome** of the recipe (The taste/texture of the cake).

Independent vs Dependent Variables

Independent variable: The variable we believe causes the change in the other variable. (e.g., Studying)

Dependent variable: The variable we are trying to predict or explain. (e.g., Test score)

🔍 Real-World Example: Predicting Customer Churn

Variable	Type	Description
Number of Customer Service Calls	**Independent** (Input)	We believe this factor **influences** churn.
Days Since Last Purchase	**Independent** (Input)	We use this data to calculate the probability.
Customer Location (Region)	**Independent** (Input)	Another factor that might affect the outcome.
Will the Customer Churn (Yes/No)?	**Dependent** (Target/Output)	This is the **final prediction** the model generates.

Relationship between variables

This model establishes a function based from the previous table:

Churn Probability = f(Calls, Last Purchase Days, Location , ……..)

The **Dependent Variable** (Churn Probability) **depends** on the **Independent Variables** (Inputs).

Models

A **model** is a **mathematical construct** (an equation or an algorithm) trained to find and quantify the relationship between past data inputs and known outcomes.

It essentially serves as a **rulebook** that takes in new, unseen data and generates a forecast (a prediction, score, or probability) for a future event.

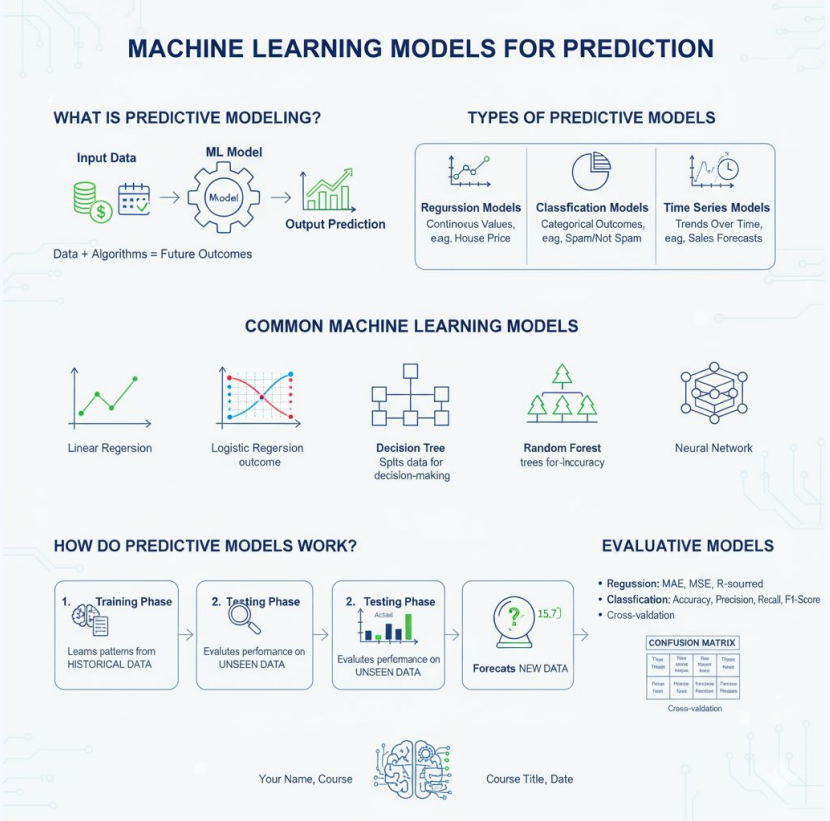
The Model's Core Function

You can think of a predictive model as a **function** that has "learned" how the world works based on historical evidence:

Model = *Trained Algorithm* → *Input Data* → *Prediction*

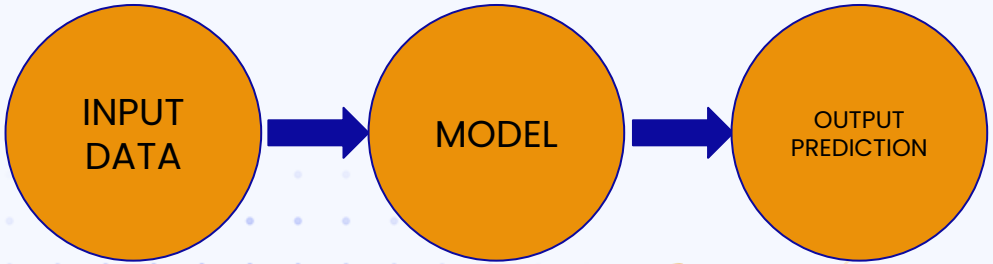
1. **Training (The Learning Phase):** A specific algorithm (like Linear Regression or a Neural Network) is fed **historical data** (the inputs) where the final outcome (the target) is already known.The algorithm iteratively adjusts its internal weights and parameters until it can accurately replicate the known outcomes from the past inputs.
2. **Forecasting (The Application Phase):** Once the model is trained, it can be given **new data** (where the outcome is currently unknown). It applies the learned mathematical rules to this new data to calculate a prediction.

Machine Learning Models for Prediction



What is Predictive Modeling ?

- Predictive modeling uses data and algorithms to forecast future outcomes.
- It is widely used in industries like finance, healthcare, marketing, and weather forecasting.
- **Goal**: Make accurate predictions based on patterns in historical data.



Three Main Types of Predictive Models

Predictive models are categorized based on the type of problem they are designed to solve:

Model Type	Purpose	Output (Dependent Variable)	Example Question
1. Regression	To predict a **continuous numerical value**.	A number (e.g., 15.7, $45,000$).	What will the **price** of the house be?
2. Classification	To predict a **category or class**.	A label or probability (e.g., **Yes**/No, **Fraud**/Not Fraud, **High** Risk).	Will this customer **default** on their loan?
3. Time Series	To predict a value that is **dependent on time** and exhibits trends or seasonality.	A number at a future date (e.g., 150 units on $10/25$).	How much **inventory** will we need next week?

Common Machine Learning Models for Prediction

1. **Linear Regression**:
   - Predicts a continuous outcome using a linear relationship.
   - Example: Predicting house prices based on size.
2. **Logistic Regression**:
   - Predicts binary outcomes (yes/no, true/false).
   - Example: Predicting if a customer will buy a product.
3. **Decision Trees**:
   - Splits data into branches for decision-making.
   - Example: Predicting loan approval.
4. **Random Forests**:
   - Ensemble of decision trees for better accuracy.
5. **Neural Networks**:
   - Complex models inspired by the human brain.
   - Example: Image recognition or stock price prediction.

How Do These Models Work ?

- **Training Phase**:
  - The model learns patterns from historical data.
  - Example: Using past sales data to identify trends.
- **Testing Phase**:
  - The model evaluates its performance on unseen data.
- **Prediction Phase**:
  - The trained model makes predictions on new data.

Evaluating Predictive Models

METRICS FOR REGRESSION MODELS

- Mean Absolute Error (MAE)
- Mean Squered Error (MSE)
- Mean Squered Error (MSE)
- R-squred

METRICS FOR CLASSIFICATION MODELS

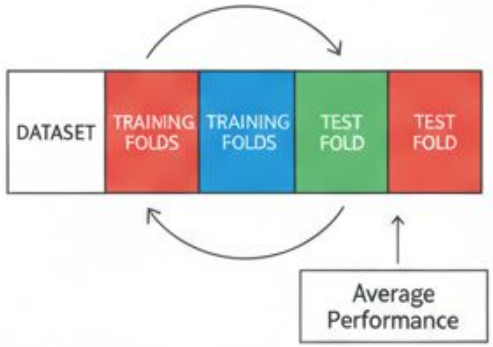
- Accuracy
- Precision
- Recall
- F1-Score

CONFUSION MATRIX

True Positive	True Positive	
False Negntive	TEST FOLD	

CROSS-VALIDATION

Ensures the model generlaizes well to unseen data.



Challenges in Predictive Modelling

Key challenges in generating predictive values include:

- **Data Quality**: Garbage in, garbage out.
- **Overfitting**: Model performs well on training data but poorly on new data.
- **Interpretability**: Some models (e.g., neural networks) are hard to interpret.
- **Ethical Concerns**: Bias in predictions, privacy issues.

Tools and Libraries for Predictive Modeling

- Python Libraries:
  - Scikit-learn: For traditional ML models.
  - TensorFlow, PyTorch: For deep learning.
  - Pandas, NumPy: For data manipulation.
- Visualization Tools: Matplotlib, Seaborn, Plotly.
- Cloud Platforms: AWS SageMaker, Google AI Platform.

PYTHON LIBRARIES

ML Models

Scrit-learn
TensrfFow
PyTorch

Data Manpulation

Pandas
Nupmp

Visualzation

Mateɾloltitt
Seab
Plotly

CLOUD PLATFORMS

aws
AWS SageMaker
Managed ML Services

Google AI Platform
Scalable Solutions

Azure
Azure Machine Learning
Seteɱming
Scalable Solutions

KEY BENEFITS

- Rapid Protopying
- Scalbility for Big Data
- Community Support
- Integration
- Deployment

Regression Analysis

What is Regression Analysis ?

- Regression analysis is a statistical technique to understand the relationship between two or more variables.

- It helps us model how one variable (the dependent variable) changes as another variable (the independent variable) changes.

Regression Analysis

Regression analysis is a statistical technique of measuring the relationship between variables. It provides the values of the dependent variable from the value of an independent variable. The main use of regression analysis is to determine the strength of predictors, forecast an effect, a trend, etc.

Regression Lines

Regression analysis creates a line (or curve) that best fits the data points.

This line represents the predicted value of the dependent variable for a given value of the independent variable.

Uses of Regression

Prediction: We can use the regression model to predict future values of the dependent variable based on new values of the independent variable. (e.g., Predicting house prices based on size)

Understanding relationships: Regression analysis helps us quantify how strongly two variables are related. Decision making: By understanding these

Why use Regression ? Practical Applications

- Prediction
  Forecast future outcomes:
  → Predict house prices based on square footage.
  → Estimate sales based on marketing spend.

- Understanding Relationships
  Quantify how strongly variables are linked:
  → Does more screen time correlate with lower sleep quality?
  → How much does temperature affect ice cream sales?

- Data-Driven Decisions
  Businesses, scientists, and policymakers use regression to:

  - Optimize pricing
  - Evaluate treatment effectiveness in healthcare
  - Personalized recommendations

What is R-Squared ?

R-squared (R^2), also called the coefficient of determination, tells you how much of the variation in your dependent variable (e.g., dengue cases) is explained by your independent variable (e.g., date/time).

1. It ranges from 0 to 1 (or 0% to 100%).

    - **$R^2 = 0$**: The model explains *none* of the variability in the data. Your predictions are no better than just guessing the average.
    - **$R^2 = 1$**: The model explains *all* the variability—perfect fit (rare in real life!).
    - **$R^2 = 0.75$**: The model explains 75% of the variation in dengue cases based on time. That's often considered quite good in real-world data!

Important Caveats

1. High $R^2 \neq$ Causation

   Just because dengue cases increase over time doesn't mean *time causes dengue*. Time might correlate with temperature, rainfall, or reporting practices.

2. R^2 Can Be Misleading

   - Adding more variables *always* increases R^2 even if they're irrelevant! (That's why sometimes use Adjusted R^2 in multiple regression.)
   - A low R^2 isn't always bad! In social or biological data (like disease trends), 30–50% can still be meaningful.

3. Always Visualize!

   Pair R^2 with a scatter plot + regression line

Bringing it to Life with Python

1. Load real-world data
2. Visualize relationships with `matplotlib` and `seaborn`
3. Build and evaluate a regression model using `scikit-learn`
4. Interpret key outputs like slope, intercept, and R-squared

Machine Learning in Python

Using Dengue Data to conduct Linear Regression

```
# Import necessary libraries
from sklearn.linear_model import LinearRegression
import numpy as np
import pandas as pd

# Convert dates to ordinal for regression
monthly_cases['Date_Ordinal'] = monthly_cases['Date'].map(pd.Timestamp.toordinal)

# Reshape data for sklearn
X = monthly_cases['Date_Ordinal'].values.reshape(-1, 1)
y = monthly_cases['Dengue_Cases'].values

# Fit the linear regression model
model = LinearRegression()
model.fit(X, y)

# Make predictions
monthly_cases['Predicted_Cases'] = model.predict(X)

# Evaluate the model: R-squared
r_squared = model.score(X, y)
print(f"The model explains {r_squared:.1%} of the variation in dengue cases.")
```

Dissecting the Code

- Step 1: Preprocessing Dates
  - Convert dates to ordinal values (numeric representation) so the model can process them.
  - Example: `2023-01-01` → `738578`.
- Step 2: Prepare Data
  - `X`: Independent variable (date ordinals).
  - `y`: Dependent variable (dengue cases).
- Step 3: Train the Model
  - Use `LinearRegression` from `sklearn` to fit the model.
- Step 4: Predict and Evaluate
  - Predict future dengue cases based on the trained model.
  - Evaluate the model's performance using the R-squared metric.

Clarification

Concept	Correlation	Linear Regression
What it measures	Strength/direction of linear relationship	Relationship for prediction and explanation
Range of output	-1 to +1	Coefficients and R^2
Symmetry	Symmetric	Asymmetric
Use case	Exploratory analysis	Predictive modeling

THANK YOU