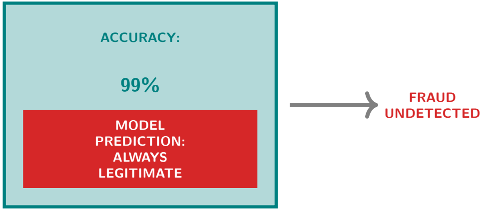
Why can a model with excellent accuracy still make terrible financial decisions?

The accuracy paradox:

- A fraud detection model reports accuracy of ninety-nine percent
- Regulators are satisfied, deployment is approved
- Within three months, millions in fraud losses occur
- Investigation reveals: the model predicted legitimate for every transaction

Why this happens:

- When fraud is one percent of transactions, predicting always legitimate yields ninety-nine percent accuracy
- The model learned to maximize the wrong metric
- Imbalanced datasets make accuracy a misleading signal

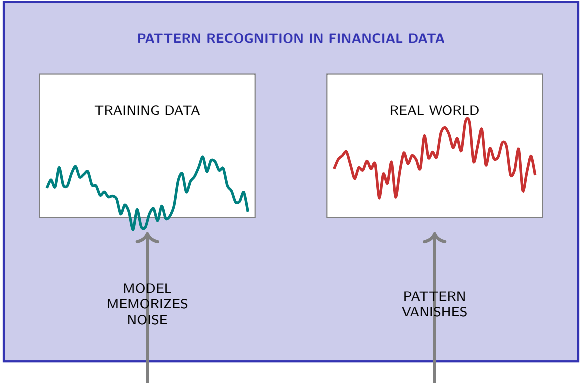


Key Insight

Accuracy measures correctness but ignores business cost. In fraud detection, catching fraud matters more than overall accuracy.

Lesson 5.1: ML Foundations for Finance — Core tension: models find patterns humans cannot see, but those patterns may be noise, bias, or measurement artifacts.

Have you ever noticed a pattern that turned out to be pure coincidence?



Reflection

Financial markets are filled with coincidental patterns. Machine learning can memorize noise in training data and mistake it for signal.

Overfitting is the primary risk in financial ML: the model learns to predict historical data perfectly but fails on new data.

What are the fundamental types of machine learning used in finance?

Supervised Learning:

- Learn from labeled examples
- Classification: predict category (fraud or legitimate, approve or deny)
- Regression: predict number (house price, revenue forecast)
- Use cases: credit scoring, fraud detection, default prediction

Unsupervised Learning:

- Learn from unlabeled data
- Clustering: group similar customers
- Anomaly detection: find unusual transactions without labeled fraud examples
- Use cases: customer segmentation, anti-money laundering

Supervised	Unsupervised
Requires labels	No labels needed
Predict specific outcome	Discover patterns
Example: credit score	Example: customer segments

Key distinction:

- Supervised answers what kind or how much
- Unsupervised finds structure without being told what to look for

Key Insight

The choice between classification and regression determines your metrics, loss function, and evaluation strategy.

Most financial ML applications use supervised learning because outcomes are well-defined and labeled data exists.

How does a decision tree learn to classify loan applicants step by step?

Tree-building algorithm:

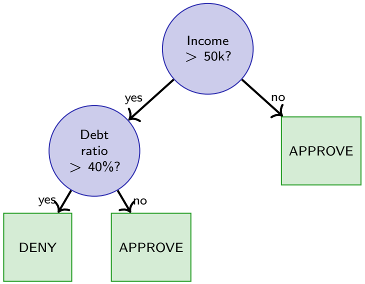
- Start with all data at the root
- Find the feature and threshold that best splits the data
- Split: income above fifty thousand or below
- Repeat recursively for each branch
- Stop when a leaf is pure or a stopping criterion is met

Strengths:

- Highly interpretable
- Handles mixed data types
- No feature scaling needed

Weaknesses:

- Prone to overfitting
- Unstable with small data changes



Key Insight

Decision trees are the building block for ensemble methods like random forest and gradient boosting.

Single trees overfit easily. Ensemble methods combine many trees to reduce variance and improve generalization.

How are the training, validation, and deployment stages of an ML pipeline structured?

Stage 1: Training

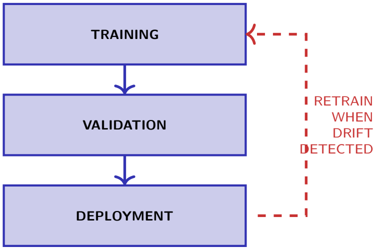
- Model learns patterns from historical data
- Fit parameters to minimize error on training set
- Risk: model memorizes training data

Stage 2: Validation

- Evaluate model on held-out data
- Tune hyperparameters using cross-validation
- Critical: use time-aware splits in finance

Stage 3: Deployment

- Deploy model to production
- Monitor performance over time
- Retrain when performance degrades



Key Insight

Financial ML is not a one-time exercise. Models must be continuously monitored and retrained as market conditions change.

Time-aware validation is critical in finance. Never train on future data to predict the past.

What happens when a model that worked perfectly in testing fails in production?

Model drift scenarios:

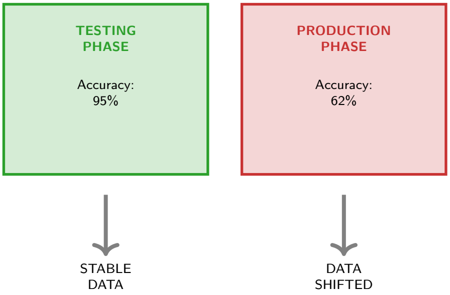
- Training data: economic growth period
- Production: recession begins
- Training patterns no longer apply
- Model performance collapses

Types of drift:

- Data drift: feature distributions change
- Concept drift: relationship between features and target changes
- Both require retraining

Prevention:

- Monitor prediction distributions
- Track performance metrics over time
- Set up automated alerts



Key Insight

Markets are non-stationary. Past performance is no guarantee of future results applies to ML models too.

Model drift is inevitable in finance. The question is not if but when your model will need retraining.

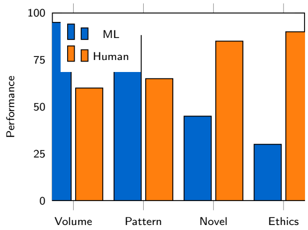
Where does machine learning outperform human judgment in finance – and where does it not?

ML outperforms humans:

- High-volume repetitive tasks
- Pattern recognition in large datasets
- Consistency over time
- Processing speed
- Example: screening thousands of loan applications per hour

Humans outperform ML:

- Novel situations outside training data
- Ethical judgment and fairness
- Explaining decisions to stakeholders
- Adapting to rare events
- Example: evaluating a unique merger scenario



Key Insight

ML and human judgment are complementary. The best systems combine both.

Machine learning excels at scale and pattern recognition. Humans excel at judgment and adaptation.

Who is helped and who is harmed when financial decisions are automated?

Who benefits:

- Borrowers with strong documented histories
- Institutions processing high volumes
- Customers seeking faster decisions
- Analysts freed from repetitive tasks

Who may be harmed:

- Applicants with non-traditional profiles
- Groups underrepresented in training data
- Those who need human judgment for context
- Workers whose jobs are automated

Fairness challenges:

- ML models can amplify historical biases
- Proxy discrimination through correlated features

Beneficiaries	At Risk
Standard profiles	Non-standard
Fast processing	Nuance needed
Cost savings	Job displacement

Regulatory concern:

- Equal Credit Opportunity Act
- Fair lending laws
- Explainability requirements

Key Insight

Automation creates winners and losers. Responsible deployment requires monitoring for disparate impact.

Fairness in ML is not just an ethical concern. It is a regulatory requirement in financial services.

Three questions to evaluate whether an ML model is ready for production in finance

Question 1: Was the model validated on out-of-time data?

- Train on past, test on future
- Never shuffle time-series data
- Simulate production conditions

Question 2: Does it perform equitably across subgroups?

- Check performance by demographic group
- Ensure no disparate impact
- Document fairness assessment

Question 3: Is there a fallback when confidence is low?

- Route uncertain cases to human review
- Set confidence thresholds
- Maintain human oversight

The Production Readiness Test:

Criterion	Pass?
Out-of-time validation	☐
Equitable performance	☐
Low-confidence fallback	☐
Performance monitoring	☐
Explainability requirement	☐
Regulatory documentation	☐

All boxes must be checked before deployment.

Key Insight

Production readiness is not just about accuracy. It requires validation, fairness, and governance.

Deploying an ML model without these checks is a regulatory risk. Model validation is mandatory in banking.

Fraud Detection Model Evaluation

A bank has deployed a fraud detection model with the following confusion matrix on a test set of one thousand transactions:

	Predicted Legitimate	**Predicted Fraud**
Actual Legitimate	920	45
Actual Fraud	5	30

Tasks:

1. Calculate precision and recall for fraud detection
2. Which type of error is more costly in this context: false positives or false negatives?
3. The model uses a threshold of point five. Recommend whether to raise or lower the threshold and explain your reasoning
4. What additional information would you need to evaluate if this model is production-ready?

Learning Goal

Apply confusion matrix metrics to a real-world financial scenario and connect technical performance to business costs.

Precision equals thirty divided by seventy-five equals forty percent. Recall equals thirty divided by thirty-five equals eighty-six percent.