

Digital Finance 3: Technology in Finance

Lesson 26: Financial Data for AI/ML

FHGR

January 3, 2026

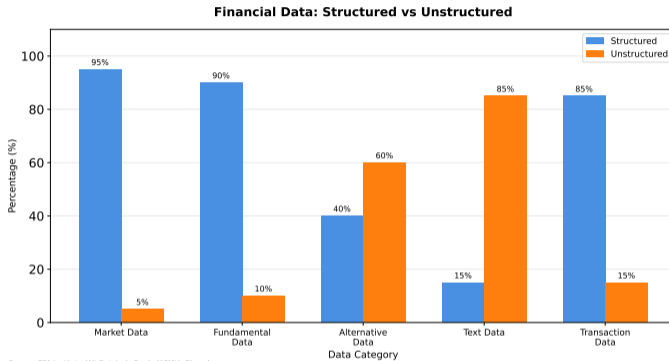
Data quality determines model quality - garbage in, garbage out.

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

- Distinguish between structured and unstructured financial data
- Identify major sources and vendors of financial data
- Understand the alternative data revolution and its applications
- Recognize data quality issues and preprocessing requirements
- Explain GDPR and privacy implications for financial ML
- Describe basic feature engineering concepts

Domain expertise in feature engineering often matters more than model complexity.

Structured vs. Unstructured Data



Modern ML: Combines both types (e.g., sentiment scores from text as features in tabular models).

Understanding different data types is essential for effective ML applications.

Traditional Financial Data Sources

Market Data:

- Stock prices (exchanges)
- Bond yields (TRACE, Bloomberg)
- Derivatives (CME, Eurex)
- FX rates (interbank, EBS)
- Crypto (Coinbase, Binance)

Fundamental Data:

- Financial statements (EDGAR, SEDAR)
- Company events (earnings, M&A)
- Economic indicators (BLS, Fed, ECB)
- Industry metrics (PMI, CPI)

Credit/Risk Data:

- Credit bureaus (Experian, Equifax, TransUnion)
- Ratings (Moody's, S&P, Fitch)
- Loan performance data
- Default histories

Major Vendors:

- \$28,320/year multi-terminal, \$31,980 single terminal (2025)^a
- Refinitiv (formerly Thomson Reuters)
- FactSet
- S&P Capital IQ
- Morningstar

^aSource: <https://www.neugroup.com/bloomberg-terminals-how-much-more-youll-pay-next-year/>

Trend: Declining costs for basic data (Yahoo Finance free), but premium data remains expensive.

Quality data is the foundation for effective machine learning models. [Source: CoinMarketCap, CoinGecko 2024]

What is Alternative Data?

- Non-traditional data sources
- Often unstructured or semi-structured
- Provides early signals
- Competitive edge (information advantage)

Categories:

- 1 **Web-scraped:** Prices, reviews, job postings
- 2 **Sensor/IoT:** Satellite, credit cards, mobile location
- 3 **Social:** Twitter sentiment, Reddit mentions
- 4 **Business:** Email receipts, app usage

Challenges: Quality control, legal/ethical concerns, data decay (alpha decay).

Example Use Cases:

- Satellite images: Count cars in parking lots (retail sales proxy)
- Credit card data: Real-time consumer spending
- Job postings: Company growth indicators
- App downloads: User engagement trends
- Shipping data: Supply chain analysis

Market Size:

- \$1.7B in 2020
- Projected \$17B by 2027
- Hedge funds are largest buyers

Financial data includes structured (prices), semi-structured (news), and unstructured (social media). [Source: McKinsey, Gartner 2024]

Satellite Imagery:

- Providers: Orbital Insight, RS Metrics
- Use: Count oil tanks, construction activity
- Example: China steel production estimates
- Frequency: Daily to weekly
- Cost: \$10k-100k+ per year

Credit Card Transactions:

- Providers: Facteus, Second Measure
- Use: Real-time revenue tracking
- Example: Restaurant chain performance
- Privacy: Aggregated, anonymized

Key Question: Does alternative data provide genuine alpha or just noise? Evidence: Mixed, diminishing returns as adoption increases (alpha decay).

Social Media Sentiment:

- Providers: RavenPack, Bloomberg sentiment
- Use: Market mood, event detection
- Example: Tweet volume predicting volatility
- Challenges: Noise, manipulation

Web Traffic:

- Providers: SimilarWeb, Alexa (discontinued)
- Use: Company engagement metrics
- Example: E-commerce site visits
- Limitation: Sample-based estimates

Case studies provide concrete evidence of technology impact and adoption patterns.

Garbage In, Garbage Out:

- ML models amplify data quality issues
- No algorithm fixes bad data
- Quality \neq Quantity (usually)

Common Data Problems:

- **Missing values:** Deletions, NaN, nulls
- **Outliers:** Errors vs. true extremes
- **Inconsistencies:** Units, formats, definitions
- **Duplicates:** Same record multiple times
- **Errors:** Typos, wrong values

Best Practice: Spend 50-80% of project time on data cleaning and validation.

Finance-Specific Issues:

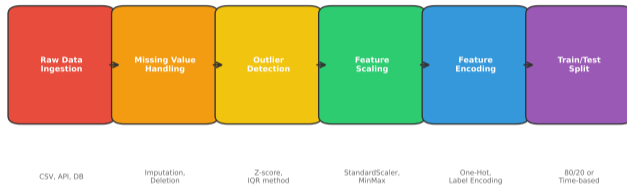
- **Survivorship bias:** Only successful firms remain
- **Look-ahead bias:** Using future information
- **Corporate actions:** Splits, dividends, mergers
- **Restatements:** Accounting changes, revisions
- **Stale data:** Delayed or infrequent updates

Impact on ML:

- Biased predictions
- Overfitting to noise
- Poor generalization
- Misleading performance metrics

Quality data is the foundation for effective machine learning models.

Financial Data Preprocessing Pipeline



Source: scikit-learn.org, pandas.pydata.org, Géron (Hands-On ML)

Automation: Modern ML pipelines use tools like Apache Airflow, Prefect for orchestration.

Systematic preprocessing is critical for model quality and reproducibility.

Why Data is Missing:

- 1 **MCAR** (Missing Completely At Random): Pure chance, no pattern
- 2 **MAR** (Missing At Random): Related to observed data
- 3 **MNAR** (Missing Not At Random): Related to unobserved value itself

Finance Example:

- MCAR: Random system glitch
- MAR: Small firms don't report segment data
- MNAR: Firms hide bad performance

Strategies:

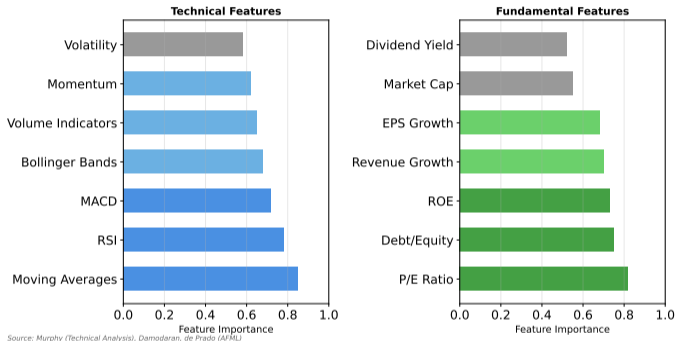
- **Deletion:** Drop rows/columns (only if $\geq 5\%$ missing, MCAR)
- **Mean/Median imputation:** Replace with average (simple, biased)
- **Forward/backward fill:** Time series (assumes persistence)
- **Model-based:** Predict missing values (KNN, regression)
- **Indicator variable:** Flag missingness as feature

Best Practice:

- Understand WHY data is missing
- Test sensitivity to imputation method
- Document assumptions

Quality data is the foundation for effective machine learning models.

Feature Engineering for Financial ML Models



Art + Science: Combines domain expertise with systematic experimentation to create predictive features.

Feature engineering often has greater impact than algorithm selection.

GDPR Key Principles (EU, 2018):

- **Lawfulness:** Legal basis for processing
- **Purpose limitation:** Specific, explicit purposes
- **Data minimization:** Collect only necessary data
- **Accuracy:** Keep data up-to-date
- **Storage limitation:** Retain only as long as needed
- **Integrity/confidentiality:** Secure processing

Individual Rights:

- Right to access
- Right to erasure (“right to be forgotten”)
- Right to explanation (Article 22)

Implications for ML:

- Consent requirements (explicit for sensitive data)
- Anonymization challenges (re-identification risk)
- Model explainability (if automated decision-making)
- Data retention policies
- Cross-border data transfers (adequacy decisions)

Other Regulations:

- CCPA (California Consumer Privacy Act)
- LGPD (Brazil)
- POPIA (South Africa)

Penalties: Up to 4% of global revenue or 20M EUR (whichever is higher).

Regulatory frameworks shape adoption patterns and industry structure. [Source: McKinsey, Gartner 2024]

Anonymization vs. Pseudonymization

Anonymization:

- Irreversible removal of identifiers
- No longer personal data under GDPR
- Techniques:
 - Aggregation
 - Noise addition (differential privacy)
 - Generalization (age → age range)
- Challenge: Re-identification risk (AOL search data, Netflix prize)

Re-identification Example: No specific 99

- ZIP code
- Gender
- Date of birth

Best Practice: Privacy by design (build privacy into system architecture from the start).

Pseudonymization:

- Replace identifiers with pseudonyms (tokens)
- Reversible with key
- Still personal data under GDPR
- Techniques:
 - Hashing
 - Encryption
 - Tokenization
- Reduces risk but doesn't eliminate it

Finance Use Cases:

- Credit scoring: Pseudonymized customer IDs
- Fraud detection: Anonymized transaction patterns
- AML: Must balance privacy with compliance

Source: Academic AI/ML literature and industry adoption studies

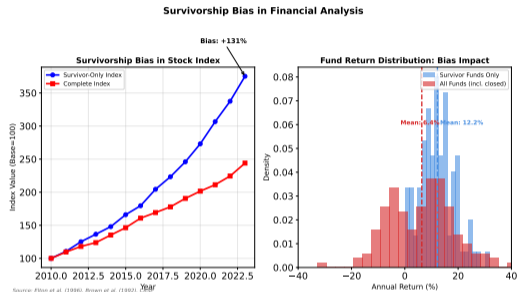
Data Bias: Types and Examples

Selection Bias:

- Sample not representative of population
- Example: Credit model trained only on approved loans

Survivorship Bias:

- Only successful entities remain in dataset
- Example: Mutual fund returns (failed funds disappear)
- Impact: Overstates historical performance



Bias in data leads to biased models—proactive detection is essential.

Unique Challenges:

- **Temporal dependence:** Observations not independent
- **Non-stationarity:** Statistical properties change over time
- **Seasonality:** Recurring patterns (quarterly earnings)
- **Trends:** Long-term drift
- **Structural breaks:** Regime changes (crises)

Cannot Use Standard ML:

- Random train/test split violates temporal order
- Cross-validation needs time-aware folds
- Risk of look-ahead bias

Proper Approach:

- **Walk-forward validation:** Train on past, test on future
- **Expanding window:** Grow training set over time
- **Rolling window:** Fixed-size window (adapts to recent data)

Feature Engineering:

- Lags ($t-1$, $t-2$, ..., $t-n$)
- Rolling statistics (moving averages, volatility)
- Date/time features (day of week, month, quarter)
- Event indicators (earnings announcement, FOMC)

Stationarity Tests:

- Augmented Dickey-Fuller (ADF)
- Differencing if non-stationary

Quality data is the foundation for effective machine learning models.

Storage Options:

- **Relational databases:** PostgreSQL, MySQL (structured data, ACID transactions)
- **NoSQL:** MongoDB, Cassandra (unstructured, scale)
- **Data warehouses:** Snowflake, Redshift (analytics)
- **Data lakes:** S3, Azure Data Lake (raw data, all formats)
- **Time series DB:** InfluxDB, TimescaleDB (high-frequency data)

File Formats:

- CSV (simple, inefficient)
- Parquet (columnar, compressed, fast)
- HDF5 (hierarchical, scientific)

Cloud vs. On-Premise:

- Cloud: Scalability, cost-effective (pay-as-you-go)
- On-prem: Control, security (for sensitive data)
- Hybrid: Regulatory compliance + flexibility

Data Governance:

- Data catalog (metadata management)
- Access control (role-based)
- Audit logs (who accessed what, when)
- Data quality monitoring
- Lineage tracking (source to destination)

Cost Considerations:

- Storage: Cheap (S3 \$0.023/GB/month)
- Compute: Expensive (processing, queries)
- Data transfer: Can be costly (egress fees)

Quality data is the foundation for effective machine learning models.

Batch Processing:

- Process large volumes periodically
- Daily, weekly, monthly updates
- Use cases:
 - Monthly credit score updates
 - Quarterly portfolio rebalancing
 - Annual financial statement analysis
- Tools: Apache Spark, Hadoop, SQL
- Pro: Efficient for large datasets
- Con: Latency (hours to days)

Lambda Architecture: Hybrid approach combining batch (accuracy) and streaming (speed) layers.

Real-Time (Streaming):

- Process data as it arrives
- Milliseconds to seconds latency
- Use cases:
 - Fraud detection (transaction monitoring)
 - Algorithmic trading (tick data)
 - AML alerts (pattern matching)
- Tools: Apache Kafka, Flink, Kinesis
- Pro: Immediate insights
- Con: Complex infrastructure, costly

Understanding the process flow is key to identifying optimization opportunities.

Bloomberg Terminal:

- Cost: \$20-25k/user/year
- Coverage: 99% of financial instruments
- Strengths: Real-time, analytics, news, messaging
- Weaknesses: Expensive, proprietary

Refinitiv (LSEG):

- Cost: \$15-30k/year
- Strengths: Historical data, fundamentals
- Eikon platform (Excel integration)

FactSet:

- Cost: \$10-15k/year
- Strengths: Quantitative analytics, screening
- Popular among asset managers

Free/Low-Cost Alternatives:

- Yahoo Finance: Free (15-20 min delay)
- Alpha Vantage: API, free tier (500 calls/day)
- Quandl: Free + premium datasets
- FRED (Federal Reserve): Economic data

Alternative Data:

- Thinknum (\$10k-50k/year)
- YipitData (\$50k+/year)
- Eagle Alpha (marketplace)

Selection Criteria:

- Coverage and accuracy
- Latency requirements
- API availability
- Cost vs. budget

Source: Academic AI/ML literature and industry adoption studies

Six Dimensions of Quality:

- 1 **Accuracy:** Values correct?
- 2 **Completeness:** All required data present?
- 3 **Consistency:** No contradictions across sources?
- 4 **Timeliness:** Data current and available when needed?
- 5 **Validity:** Conforms to format, range, type?
- 6 **Uniqueness:** No duplicates?

Measurement:

- Accuracy: % records passing validation
- Completeness: % non-null values
- Consistency: % cross-checks passing

Quality Assurance Process:

- 1 Define quality rules
- 2 Automated checks (scripts)
- 3 Exception handling
- 4 Root cause analysis
- 5 Continuous monitoring
- 6 Feedback loop to data sources

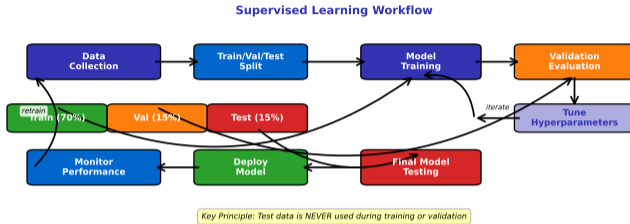
Tools:

- Great Expectations (Python library)
- dbt (data build tool)
- Custom SQL checks
- Dashboards (Tableau, Power BI)

Best Practice: Shift-left testing (validate early in pipeline, not at model training).

Quality data is the foundation for effective machine learning models.

Supervised Learning Workflow

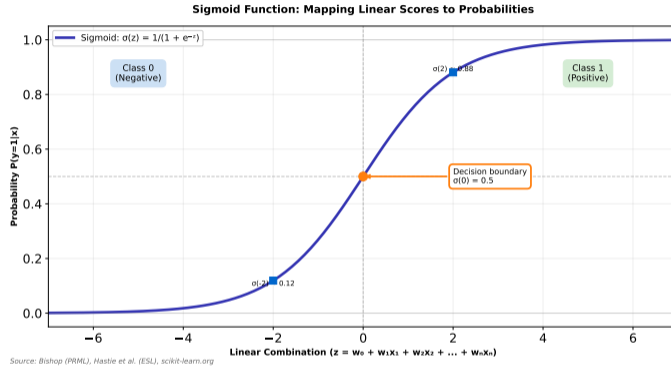


Typical Split: 70% Train, 15% Validation, 15% Test

Source: Hastie et al. (ESL), scikit-learn, Google ML Course

The standard pipeline from data to deployment requires careful validation at each stage.

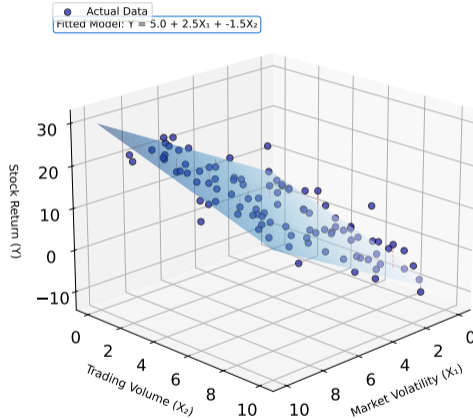
Logistic Function and Sigmoid Curve



The sigmoid function maps any real number to a probability between 0 and 1.

Multiple Regression Visualization

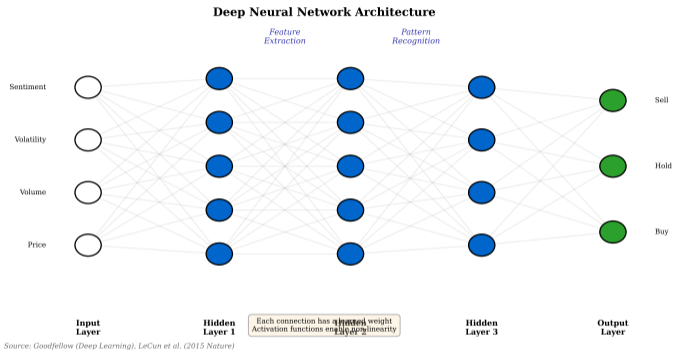
Multiple Linear Regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$



Source: Wooldridge (Econometrics), Hastie et al. (ESL), scikit-learn

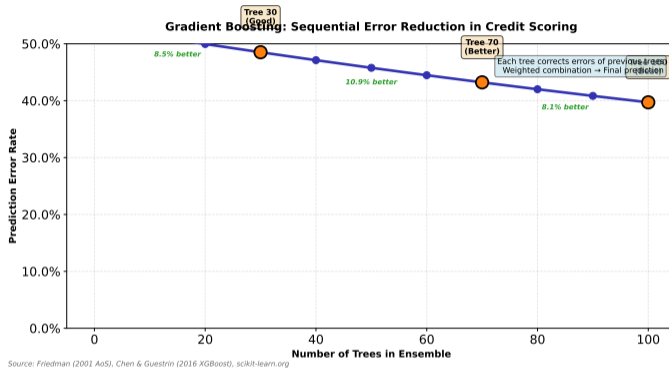
Source: Academic AI/ML literature and industry adoption studies

Deep Learning Architecture



Deep neural networks learn hierarchical representations through multiple layers.

Gradient Boosting Process



Gradient boosting builds models sequentially, each correcting errors of the previous.

Data Types:

- Structured (tables) vs. unstructured (text, images)
- Traditional (prices, financials) vs. alternative (satellite, social)
- Real-time (streaming) vs. batch (periodic)

Quality is Critical:

- GIGO principle applies
- 50-80% of effort on data work
- Bias detection and mitigation
- Time series requires special care

Regulations Matter:

- GDPR and privacy laws
- Anonymization challenges
- Right to explanation
- Compliance $\hat{=}$ model performance

Practical Considerations:

- Cost-benefit of data vendors
- Infrastructure (cloud, storage)
- Feature engineering creativity
- Continuous monitoring

Financial data includes structured (prices), semi-structured (news), and unstructured (social media).

Lesson 27: Supervised Learning - Regression

Topics to be covered:

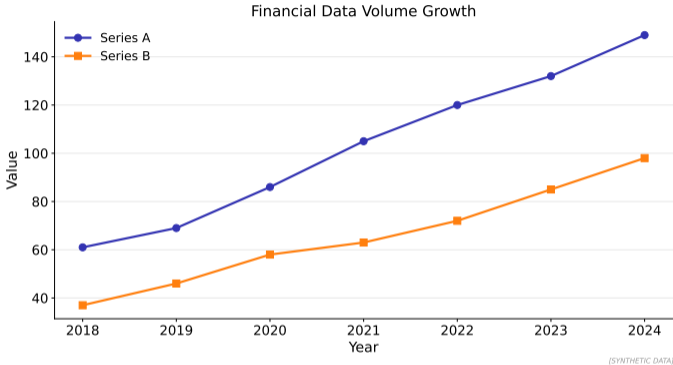
- Features, labels, and training data
- Simple and multiple linear regression
- Coefficients interpretation
- R-squared and evaluation metrics
- Overfitting and regularization (Ridge, Lasso)
- Applications: Stock return prediction, pricing models

Preparation:

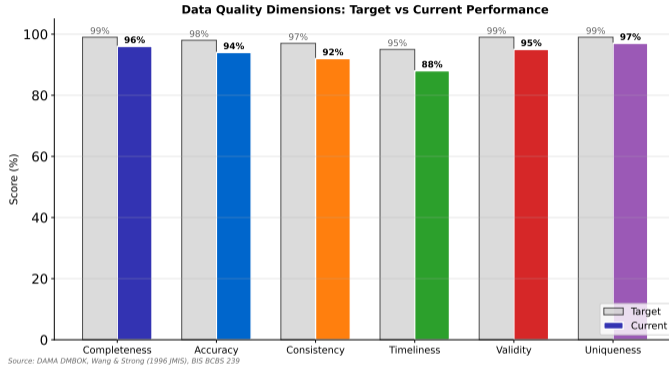
- Review basic linear algebra (vectors, matrices)
- Recall correlation and covariance concepts
- Think: What financial problems involve predicting continuous values?

Linear regression provides interpretable baselines before trying complex models.

Financial Data Volume Growth



Data volume growth drives need for advanced analytics capabilities.



Quality dimensions ensure reliable model inputs.