Quiz: Feature Engineering

AI-Based Detection of Hedge Fund Fraud

Joerg Osterrieder

Zurich University of Applied Sciences (ZHAW)

2025

Question 1

What is the typical feature space dimensionality for fraud detection?

a) 10–25
b) 25–50
c) 50–200
d) 500–1000

Question 1

What is the typical feature space dimensionality for fraud detection?

a) 10–25
b) 25–50
c) 50–200
d) 500–1000

Answer

c) 50–200

Most fraud detection systems use 50–200 engineered features, combining return statistics (mean, volatility, skewness), Benford's law digits, autocorrelation measures, and network metrics. Higher dimensions risk overfitting.

Appendix A1: Feature Engineering

Question 2

What minimum return history is needed for reliable feature engineering?

a) 6–12 months
b) 24–36 months
c) 48–60 months
d) 72–84 months

Question 2

What minimum return history is needed for reliable feature engineering?

a) 6–12 months
b) 24–36 months
c) 48–60 months
d) 72–84 months

Answer

b) 24–36 months

Reliable statistical features require at least 24–36 months of return history. Shorter windows reduce statistical power for autocorrelation, volatility clustering, and anomaly detection. This creates the "cold-start" problem.

Appendix A1: Data Requirements

Question 3

What is Benford's law first-digit frequency for $d = 1$?

a) $\sim$30.1%
b) $\sim$17.6%
c) $\sim$11.1%
d) $\sim$25.0%

Question 3

What is Benford's law first-digit frequency for $d = 1$?

a) $\sim$30.1%
b) $\sim$17.6%
c) $\sim$11.1%
d) $\sim$25.0%

Answer

a) $\sim$30.1%

Benford's law predicts that the first digit $d = 1$ appears with frequency $\log_{10}(1 + 1/1) \approx 30.1\%$. Significant deviations in reported returns suggest manipulation. Fraudsters often under-report digit 1.

Appendix A1: Benford's Law

Question 4

Which method is used for chi-squared Benford testing?

a) t-test on return means
b) Kolmogorov–Smirnov test
c) Compare observed vs expected digit frequencies
d) ANOVA across periods

Question 4

Which method is used for chi-squared Benford testing?

a) t-test on return means
b) Kolmogorov–Smirnov test
c) Compare observed vs expected digit frequencies
d) ANOVA across periods

Answer

c) Compare observed vs expected digit frequencies

The chi-squared test compares observed first-digit frequencies from reported returns against Benford's theoretical distribution. High χ^2 values indicate anomalous digit patterns consistent with fabrication.

Appendix A1: Statistical Testing

Question 5

Why is PCA less common for feature reduction in fraud detection?

a) Too slow computationally
b) Requires too much data
c) Cannot handle categorical features
d) Loss of interpretability

Question 5

Why is PCA less common for feature reduction in fraud detection?

a) Too slow computationally
b) Requires too much data
c) Cannot handle categorical features
d) Loss of interpretability

Answer

d) Loss of interpretability

Principal Component Analysis (PCA) reduces dimensionality but creates opaque linear combinations. Regulators require interpretable features (e.g., “serial correlation”, “Sharpe ratio”) for audit trails and legal justification.

Appendix A1: Dimensionality Reduction