Statistics and Data Analysis

Unit 05 – Lecture 01: Feature Selection, Engineering and Dimensionality Reduction (Intro)

Tofik Ali

School of Computer Science, UPES Dehradun

February 17, 2026

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Overview Why Features Selection vs Reduction Exercises Demo Summary

Agenda

1 Overview

2 Why Features

3 Selection vs Reduction

4 Exercises

5 Demo

6 Summary

Learning Outcomes

- Differentiate feature selection vs dimensionality reduction

Learning Outcomes

- Differentiate feature selection vs dimensionality reduction
- Explain why too many features can hurt (overfitting, cost)

Learning Outcomes

- Differentiate feature selection vs dimensionality reduction
- Explain why too many features can hurt (overfitting, cost)
- Describe a simple feature engineering pipeline

Learning Outcomes

- Differentiate feature selection vs dimensionality reduction
- Explain why too many features can hurt (overfitting, cost)
- Describe a simple feature engineering pipeline
- Identify target leakage in engineered features

Why Features: Key Points

- Features are how models see data

Why Features: Key Points

- Features are how models see data
- Goal: represent signal and reduce noise

Why Features: Key Points

- Features are how models see data
- Goal: represent signal and reduce noise
- Bad features -¿ bad models

Selection vs Reduction: Key Points

- Selection keeps a subset of original features

Selection vs Reduction: Key Points

- Selection keeps a subset of original features
- Reduction creates new components (e.g., PCA)

Selection vs Reduction: Key Points

- Selection keeps a subset of original features
- Reduction creates new components (e.g., PCA)
- Validate choices using CV

Exercise 1: Selection or reduction

Dropping 30 out of 100 features is selection or reduction?

Solution 1

- Feature selection (subset).

Exercise 2: Leakage

Is using final exam score to predict final grade leakage?

Solution 2

- Yes; it contains future/target information.

Exercise 3: Engineering example

Give one time-based engineered feature.

Solution 3

- Day-of-week, month, time-since-last-event, rolling average, etc.

Mini Demo (Python)

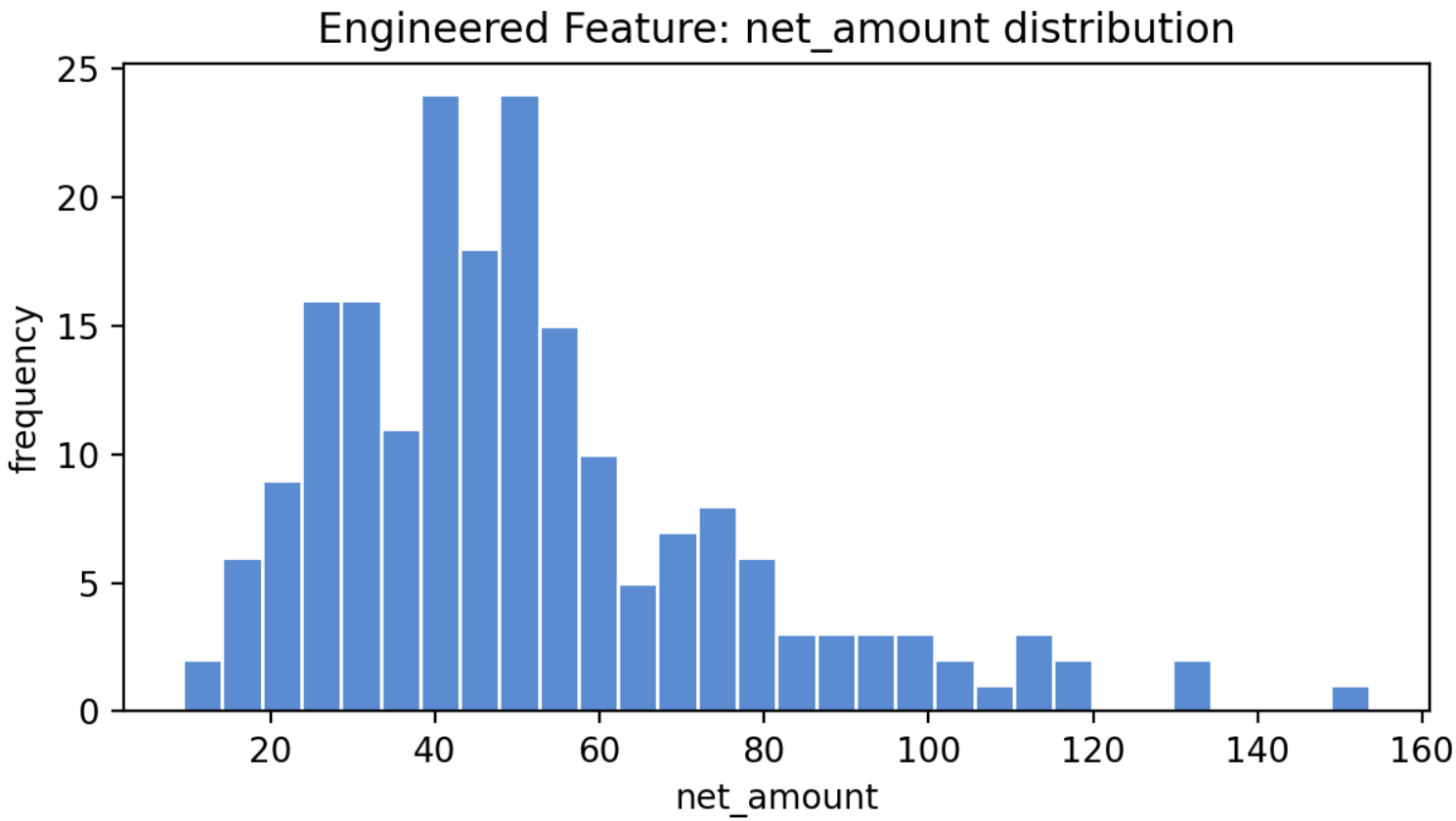
Run from the lecture folder:

```
python demo/demo.py
```

Outputs:

- `images/demo.png`
- `data/results.txt`

Demo Output (Example)



Summary

- Key definitions and the main formula.

Summary

- Key definitions and the main formula.
- How to interpret results in context.

Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Why can adding more features sometimes reduce test accuracy?