Statistics and Data Analysis
Unit 05 – Lecture 06 Notes

Tofik Ali

February 14, 2026

Topic

Interactions, aggregations, time features, and leakage avoidance.

Learning Outcomes

- Create interaction features when meaningful
- Create aggregation features from transactional data
- Engineer time-based features (lags/rolling)
- Avoid leakage and look-ahead bias

Detailed Notes

These notes are designed to be read alongside the slides. They expand each slide bullet into plain-language explanations, small worked examples, and common pitfalls. When a formula appears, emphasize (1) what each symbol means, (2) the assumptions needed to use it, and (3) how to interpret the final number in the problem context.

Interactions

- Products and ratios capture combined effects
- Use domain knowledge
- Validate with CV

Aggregations

- Per-user totals/means/counts
- Rolling windows (last 7/30 days)
- Avoid using future data

Exercises (with Solutions)

Exercise 1: Interaction

Give one interaction feature for house price.

Solution

- size_m2 * location_score (example).

Exercise 2: Aggregation

Name one per-user aggregation for churn prediction.

Solution

- days_since_last_purchase (example).

Exercise 3: Leakage

Is using next-30-days spend to predict churn today leakage?

Solution

- Yes; it uses future info.

Exit Question

How does cross-validation help detect whether engineered features overfit?

Demo (Python)

Run from the lecture folder:

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

Output files:

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

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley.
- Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, Cengage.
- McKinney, W. *Python for Data Analysis*, O'Reilly.