A Report on the UCI Bike Sharing Project

Exploratory Analysis, Regression of Counts, and Hour Classification

Dominik Sieber

October 12, 2025

1 Task at hand

Source. UCI Machine Learning Repository: Bike Sharing Dataset (ID 275). We use the hourly data for 2011–2012 (Washington D.C.), comprising calendar variables, meteorology, categorical weather, and the total number of rentals `cnt`.

The exercise consists of two tasks:

Task 1: Regression of counts Build a linear baseline to predict `cnt` from a *minimal* and well-justified feature set. Compare to a blind mean baseline and report RMSE/MAE/R^2.

Task 2: Classification of hour Train a multinomial logistic regression to predict the hour $y \in \{0, \ldots, 23\}$. Use a compact feature set, normalise any direct use of `cnt` (if included), compare against uniform (1/24) and majority-hour priors, and report accuracy, macro-F1, log loss, and mutual information.

To ensure we pick the best features for these task we do *EDA first* to understand the data structure.

2 Exploratory Data Analysis

2.1 Temporal structure

Figure 1 displays the hourly time series. Counts increase from winter to summer with year-over-year trend.

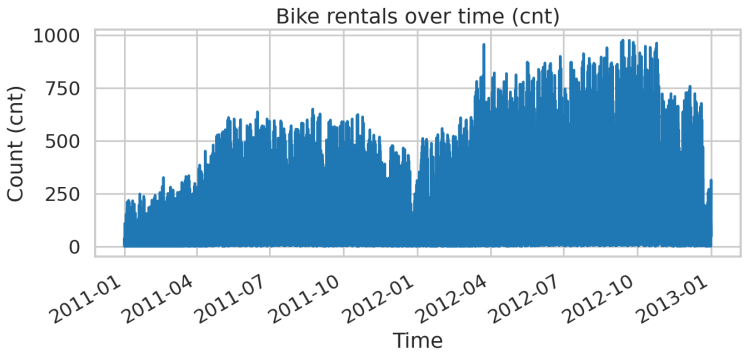
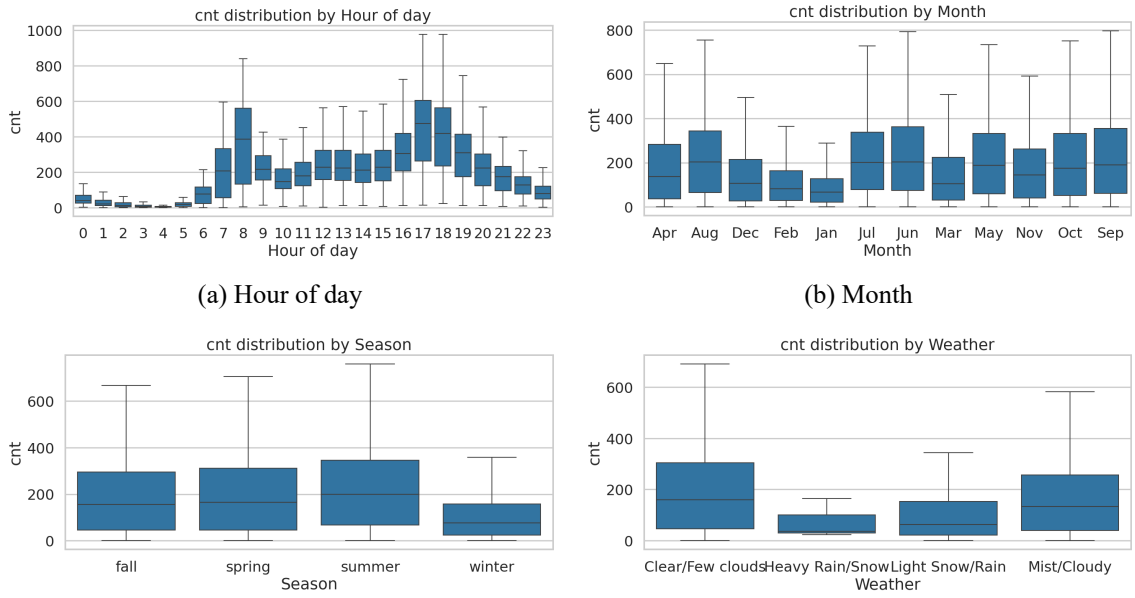


Figure 1: Hourly rentals over time (2011–2012). Seasonal structure and upward trend. strong daily periodicity.

2.2 Distributions across categorical drivers

Figure 2 summarises `cnt` by hour, month, season, and weather. Hour shows a bi-modal pattern with commute peaks at 07–08 and 17–18, minima per hour hit zero at any hour. Warmer months/seasons have higher counts, heavy rain/snow suppresses demand most.



(a) Hour of day

(b) Month

(c) Season

(d) Weather

Figure 2: Count distributions by categorical factors.

2.3 Categorical–numeric association and numeric correlations

The correlation ratio η (category $\rightarrow$ numeric) ranks hour as most informative, followed by month/season and then weather (Figure 3). For numeric features, temperature proxies correlate most with counts ($r \approx 0.4$–0.5), windspeed more weakly (≈ 0.2). The relation of counts to apparent temperature is non-linear and dome-shaped (Figure 4).

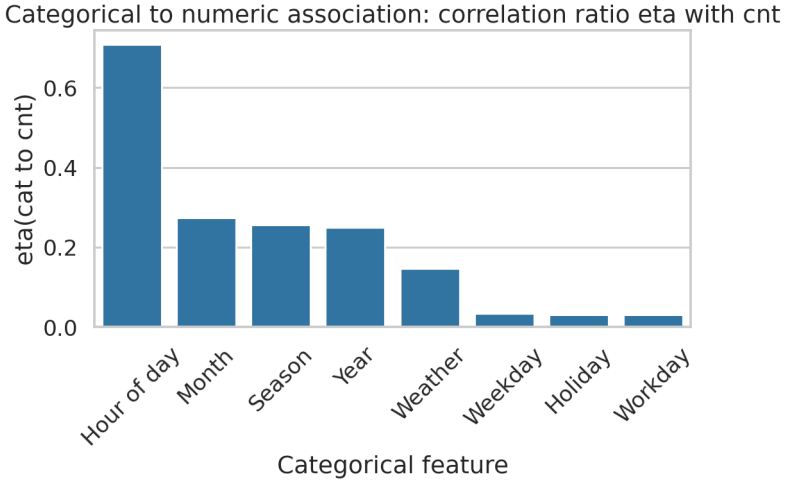
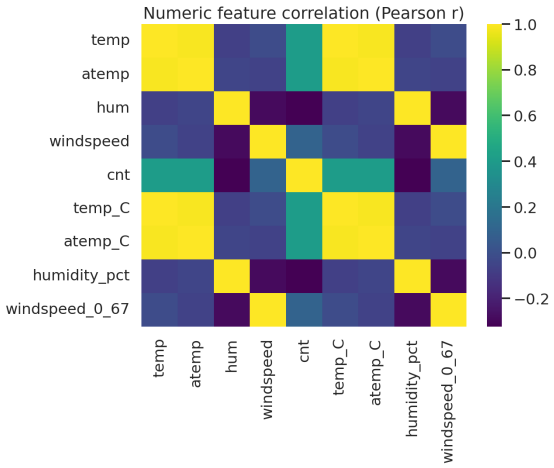
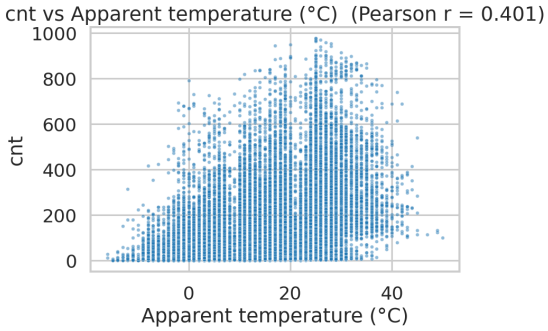


Figure 3: Correlation ratio η of categorical features with `cnt`. Hour dominates, month/season next, weather contributes, weekday/holiday/working day weak marginally.



(a) Pearson correlation matrix



(b) `cnt` vs. apparent temperature (°C)

Figure 4: Numeric associations: temperature is most predictive, with clear non-linearity.

3 Task 1: Linear regression of `cnt`

Implementation Because the data are time-ordered, we split *chronologically* into train, validation, and test blocks to avoid leakage from seasonality. The model is standard linear regression with ℓ_2 regularisation (ridge), as in the lecture notes: training minimises squared error on the chosen target.

Feature selection by greedy ablation A custom routine evaluates candidate feature subsets by fitting ridge on train and selecting by validation RMSE. In our run, the first large gain comes from adding a one-hot encoding of *hour-of-day*, followed by a quadratic temperature proxy (`atemp`, `atemp`^{2}). Further additions (weather bins, seasonal dummies, extra Fourier hour terms) did not improve validation RMSE appreciably (see Figure 5).

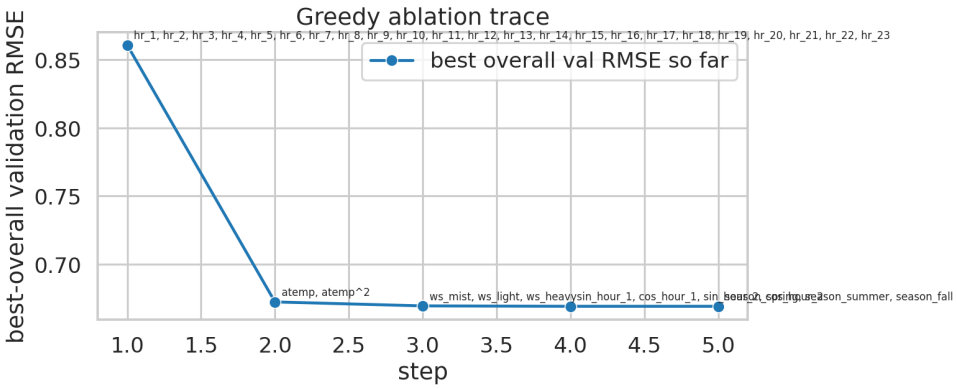


Figure 5: Greedy ablation trace (validation RMSE). Step 1 selects hour dummies, step 2 adds `atemp` and `atemp`^{2}. Later additions yield negligible improvement.

Baselines. Figure 6 compares test RMSE across: (i) mean-only baseline, (ii) an hour-only baseline (time prior), and (iii) ridge with the selected features. Ridge substantially beats the *mean* baseline, but it does *not* outperform the hour-only baseline for this feature set. This points to bias from the linear specification / limited information density, not variance/overfitting.

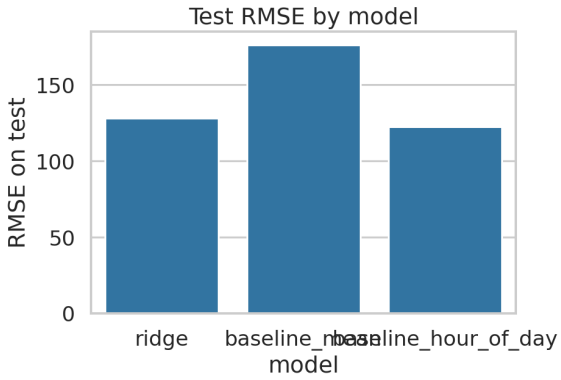


Figure 6: Test RMSE by model. Ridge $>$ mean baseline, hour-only baseline remains competitive and slightly better than ridge in this configuration.

Regularisation sweep. A wide grid over λ shows a flat validation RMSE valley and degradation only for very large λ (Figure 7). Hence regularisation is not the limiting factor, the gap is driven by model bias and feature information.

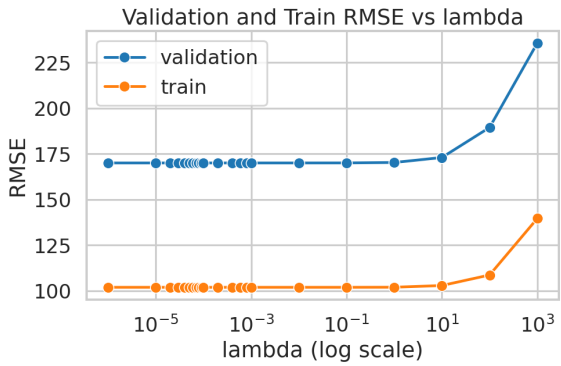


Figure 7: Train/validation RMSE vs. ridge λ (log scale). Flat valley indicates regularisation is not the driver of the gap to the hour-only baseline.

Residual analysis. Define residuals as $r = y - \hat{y}$ (positive $\Rightarrow$ under-prediction). Averaging r by hour shows large positive residuals at commute peaks, especially on the test block (Figure 8). The validation block displays two peaks but still with positive means (Figure 9). The model captures the daily *periodicity* yet systematically *underestimates* peak magnitudes.

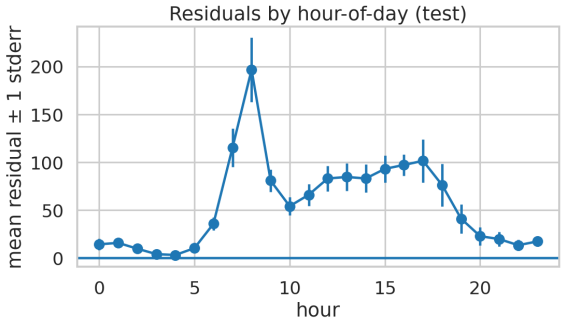


Figure 8: Mean residual $\pm$ 1 by hour (test). Positive spikes at commute hours indicate under-prediction at peaks.

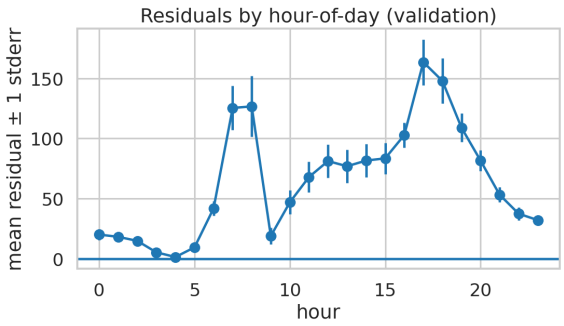


Figure 9: Mean residual $\pm$ 1 s.e.m. by hour (validation). Both peaks appear but remain under-predicted.

Temporal fit quality. The overlaid time series on the test block (Figure 10) confirms that ridge reproduces the intra-day oscillation but consistently under-represents peak counts.

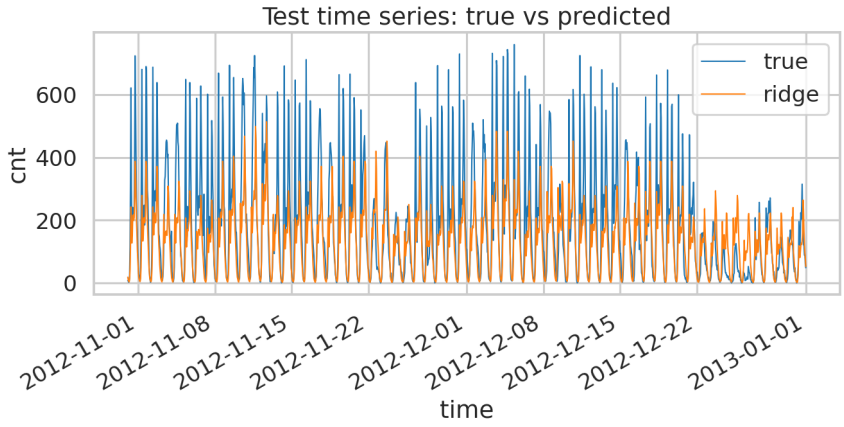


Figure 10: Test period, true vs. ridge predictions. Periodicity captured, peak amplitudes routinely under-predicted.

4 Task 2: Hour classification

Setup. We frame hour prediction as multinomial logistic regression with chronological train/validation/test splits (as in Task 1). The most informative predictor for hour is the total count `cnt`. To stabilise optimisation and make logits/gradients comparable across features, we scale counts to the unit interval,

$$\tilde{c} = \frac{\texttt{cnt} - c_{\min}}{c_{\max} - c_{\min}} \in [0, 1], \tag{1}$$

which (i) prevents a single large-scale feature from dominating the softmax, (ii) yields well-scaled gradients early in training, and (iii) improves numerical conditioning. Candidate features additionally include apparent temperature (`atemp`, $\texttt{atemp}^2$) and weather categories as well as others (but not selected in the greedy ablation).

Greedy ablation. We repeat the greedy subset selection used for regression, now maximising validation accuracy. As shown in Figure 11, using $\tilde{c}$ alone reaches $\approx$ 13% validation accuracy. Adding `atemp` (and its square) and coarse weather indicators nudges accuracy to $\approx$ 14–15%

(on train, validation remains near $\approx$ 13–14%). This clearly exceeds uniform random guessing ($1/24 \approx 4.17\%$), but remains far from reliable hour identification. The minimal useful set is therefore $\{\tilde{c}, \texttt{atemp}, \texttt{weathersit}\}$.

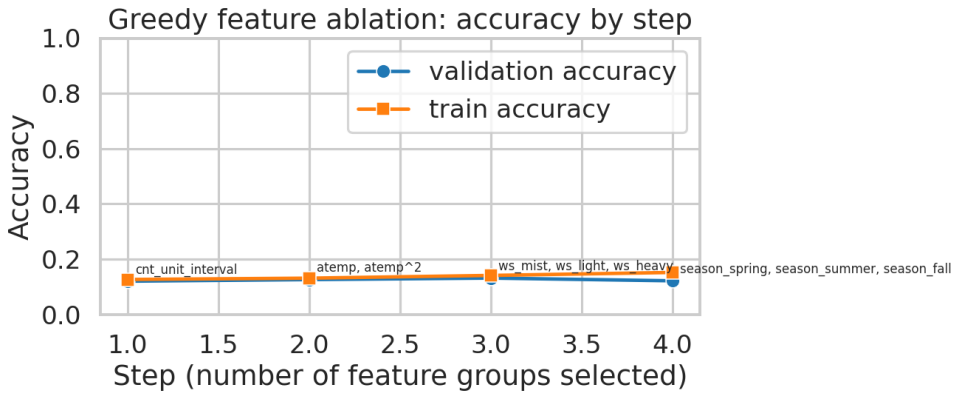


Figure 11: Greedy ablation for hour classification. Counts (unit interval) dominate, temperature and weather add small gains.

Per-hour behaviour. Figure 12 plots accuracy by hour for each split with class supports in the background. Night hours (low usage) and the evening peak are comparatively easier. Most mid-day hours remain hard. The test block shows good performance around 4–5 and 8, but a weaker spike at 17 compared to train/validation. This asymmetry reflects distribution shift and intrinsic ambiguity: many hours share overlapping count ranges.

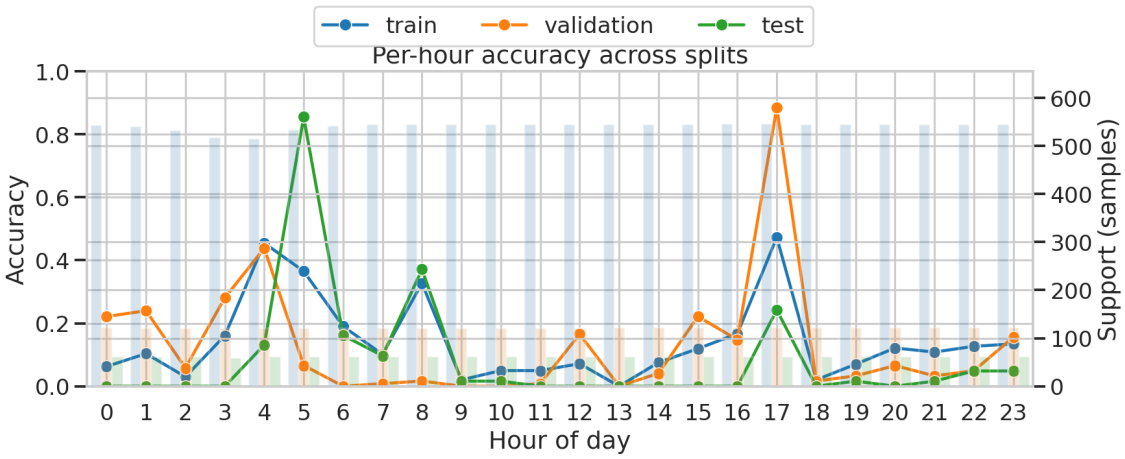


Figure 12: Per-hour accuracy across splits with class supports. Peaks around typical commute or very low-usage hours, mid-day poorly separated.

Optimisation diagnostics. Training/validation loss and in-sample accuracy flatten early (Figure 13), indicating that longer training is not the bottleneck. With the chosen features the model reaches its capacity plateau, the limiting factor is class overlap in feature space.

Confidence analysis. Figure 14 shows the predicted probability assigned to the *true* hour, stratified by correctness. Correct predictions have higher assigned probabilities than incorrect ones, but both

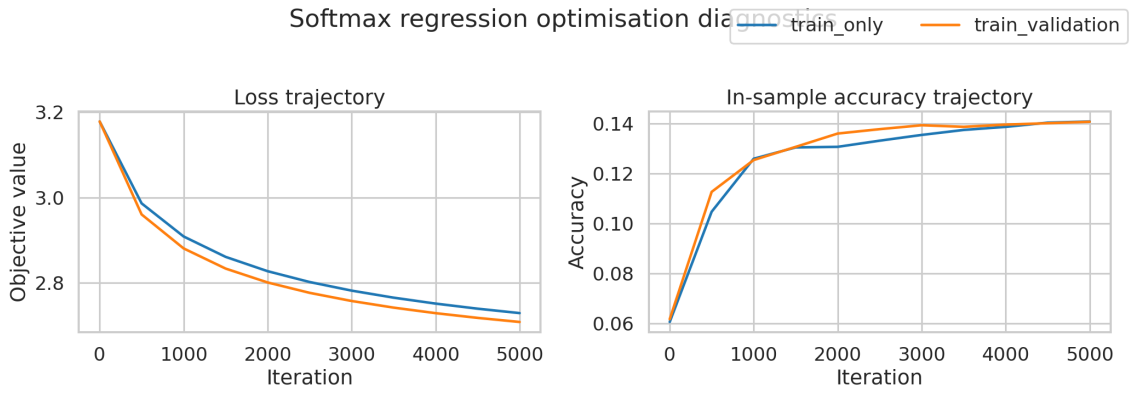


Figure 13: Softmax regression optimisation: objective and accuracy trajectories. Both curves plateau, more iterations do not really help.

concentrate at low values (< 0.15). This means the classifier is seldom confident because the same $(\tilde{c}, \texttt{atemp}, \texttt{weathersit})$ combinations occur at many hours.

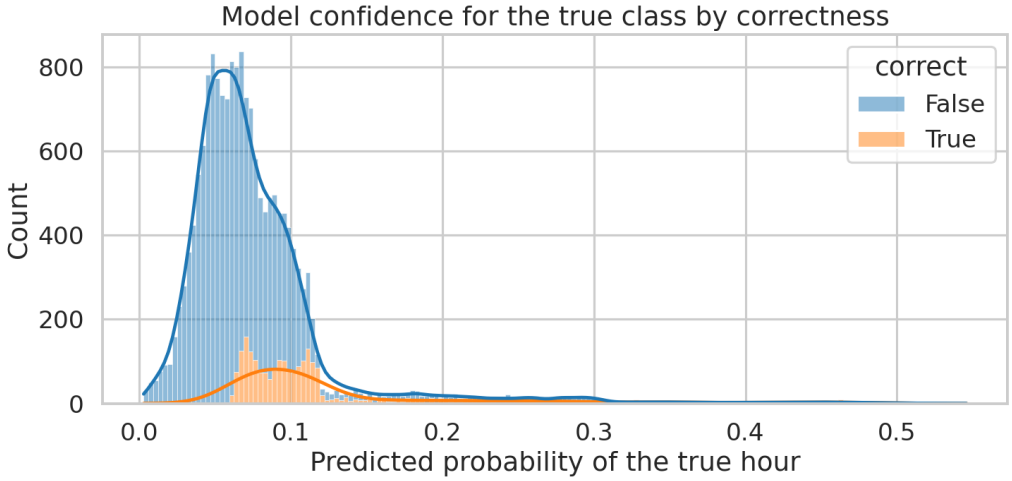
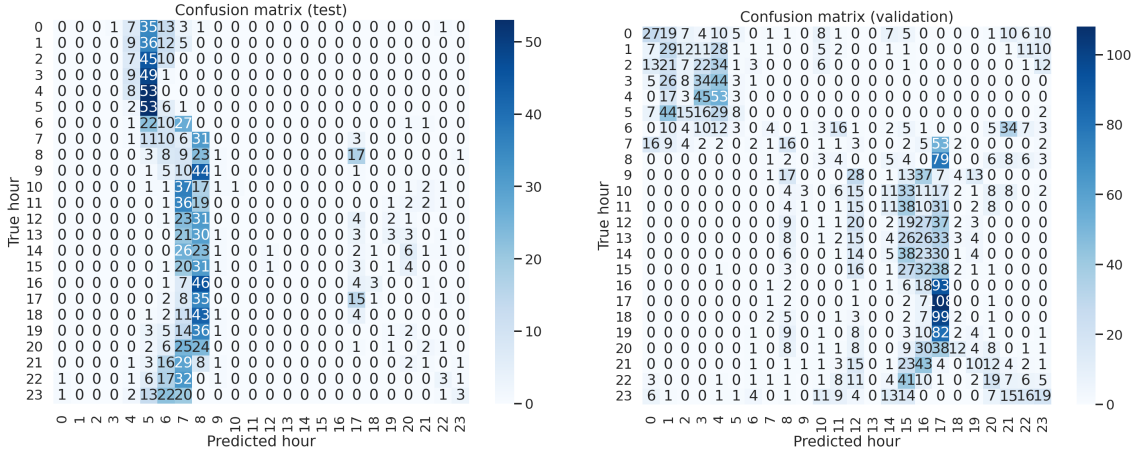


Figure 14: Histogram of $p_\theta(y_{\text{true}} \mid x)$. Even correct predictions rarely exceed moderate confidence, feature overlap limits separability.

Confusion structure. The ideal confusion matrix is diagonal. Our matrices (Figure 15) exhibit strong mass around a few adjacent hours and along columns for popular classes. Misclassifications cluster near peak hours where neighbouring classes are indistinguishable given the features. Differences between validation and test suggest mild temporal shift of peak positions/heights across blocks.

Conclusion. Although daily periodicity, season, and weather shape the *distribution* of counts, many distinct hours can realise similar $(\tilde{c}, \texttt{atemp}, \texttt{weathersit})$. Zero counts occur at almost any hour, peaks shift in height/width across days and seasons. Without stronger temporal context (like lagged counts, explicit time-of-day priors, or sequential models), discriminating between neighbouring hours is close to random for large portions of the day. The observed accuracies and low confidences are consistent with this information-theoretic limitation.



(a) Test

(b) Validation

Figure 15: Confusion matrices. Off-diagonal mass concentrates around commute bands and adjacent hours.

References

[1] UCI Machine Learning Repository. Bike Sharing Dataset (ID 275). `https://archive-beta.ics.uci.edu/dataset/275/bike+sharing+dataset`.

[2] Lecture notes from Markus Schmitt for Modern Machine Learning, 2025.