A Report on the UCI Bike Sharing Project

Exploratory Analysis, Regression of Counts, and Hour Classification

Dominik Sieber

October 26, 2025

Assessment rubric (25 points total)

Block	Pts	What must be shown / explained to earn points
Task 1: Regression	**10**	*Chronological split* with rationale (seasonality/leakage) **(2)**. *Baselines model* (random guessing) overlay true vs. prediction with per-hour averages and discussion that peaks are under-predicted **(2)**. *Feature exploration* (no auto ablation required) concluding hour is dominant, others minor - Figuring out that the hour should be encoded via sine and/or cosine **(3)**. *Regularisation test* (ridge) and argument why it has negligible effect here (the model is the problem -> bias not variance) **(1)**. *Periodicity discussion* of the data and how this impacts the model **(2)**.
Task 2: Classification	**10**	*Feature exploration* showing why (normalised) count is most informative and how normalisation to $[0, 1]$ is done **(2)**. Showing if and how other features impact the classification **(2)**. *Accuracy across splits* with class supports and why commute/peak hours are easier **(2)**. *Optimisation diagnostics* (loss trajectory, in-sample accuracy) **(2)** and explaining why it can't go lower. *Mutual information* estimate/discussion from $P_\theta(y \mid x)$ **(2)**.
Presentation	**5**	Clear, concise figures with informative captions.

Note. Feature exploration should be reasoned trial-and-error. An automatic exhaustive ablation search is *not* required.

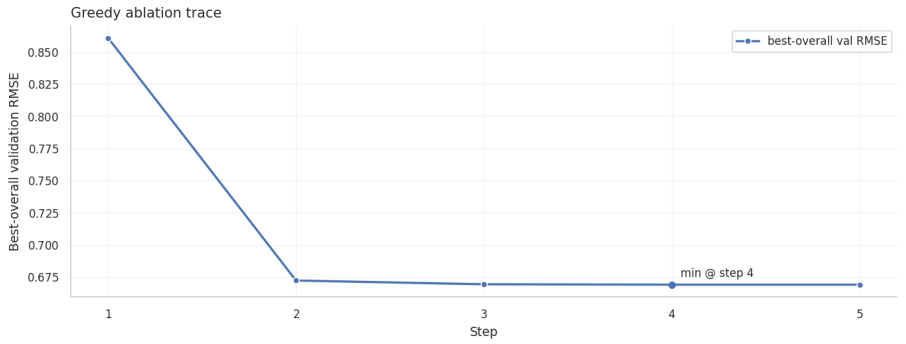
1 Task 1: Linear regression of `cnt`

Data split and model. 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 (total count `cnt`).

Points

[2 pt] Chronological split used and justified.

Feature exploration. Guided by periodicity, we experiment with: one-hot *hour-of-day*, apparent temperature (`atemp`, optionally `atemp`^{2}), coarse weather categories, and calendar dummies. The largest gain comes from adding the hour representation second to the temperature, as the latter influences bike loaning the most. additional covariates add comparatively little signal for linear models on chronologically held-out data (see Figure 1 where automatic ablation selection shows that not many features are needed to get good results).



Step	Chosen group
1	hr_1, hr_2, hr_3, hr_4, hr_5, hr_6, hr_7, hr_8, hr_9, hr_10, hr_11, hr_1
2	atemp, atemp^2
3	ws_mist, ws_light, ws_heavy
4	sin_hour_1, cos_hour_1, sin_hour_2, cos_hour_2
5	season_spring, season_summer, season_fall

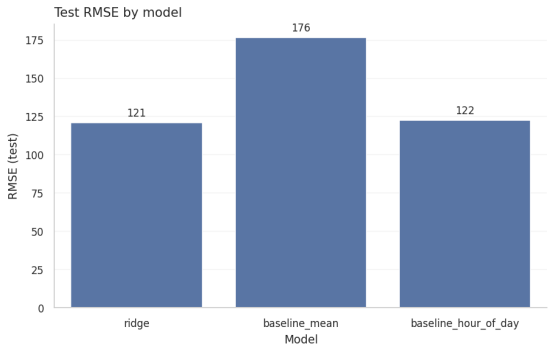
Figure 1: Greedy ablation trace (validation RMSE). Step 1 selects hour dummies, step 2 adds `atemp` and `atemp`^{2}. Later additions yield negligible improvement.

Points

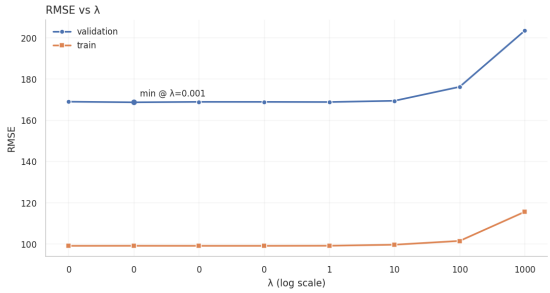
[3 pt] Explain *why* hour is dominant (periodicity). Justify limited value from other features.

Baselines and what the model really predicts. Figure 2a 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.

A wide grid over λ shows a flat validation RMSE valley and degradation only for very large λ (Figure 2b). Hence regularisation is not the limiting factor, the gap is driven by model bias and feature information.



(a) Test RMSE by model. Ridge > mean baseline, hour-only baseline remains competitive and slightly better than ridge in this configuration.



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

Points

[1 pt] Show/train–val curves vs. λ and argue why regularisation is not crucial here.

The overlaid time series on the test block (Figure 3) confirms that ridge reproduces the intra-day oscillation but consistently under-represents peak counts.

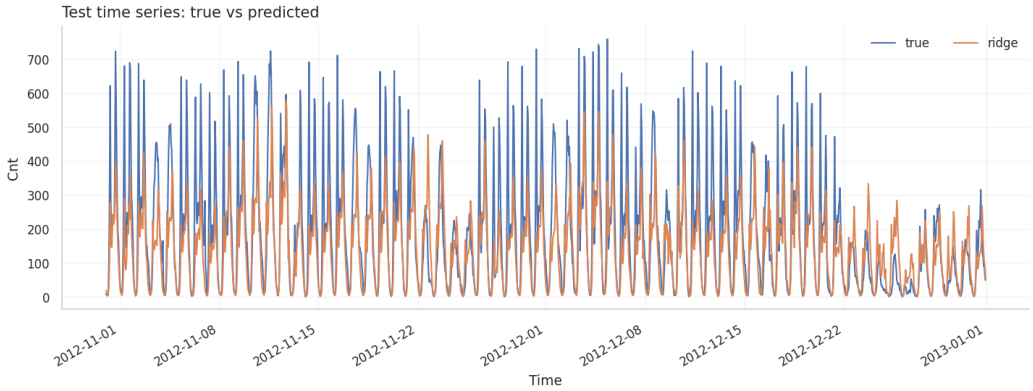


Figure 3: Test period, true vs. ridge predictions. Periodicity captured, peak amplitudes routinely under-predicted.

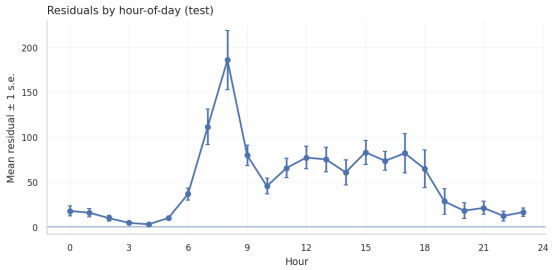
Points

[2 pt] Provide baseline numbers/plots in comparison with the model.

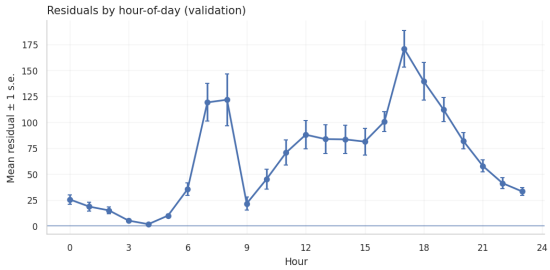
Residual analysis. Define residuals as $r = y - \hat{y}$ (positive $\Rightarrow$ under-prediction). Averaging r by hour shows large positive residuals at commute peaks, especially on the test block (Figure 4a). The validation block displays two peaks but still with positive means (Figure 4b). The model captures the daily *periodicity* yet systematically *underestimates* peak magnitudes.

Points

[2 pt] Show residuals by hour (or rather see which in average at which hours the counts are best predicted) and discuss periodicity and peak bias.



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



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

2 Task 2: Hour classification

Setup and feature scaling. 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 or when tried by trial-and-error should not be as good as a predictor).

Points

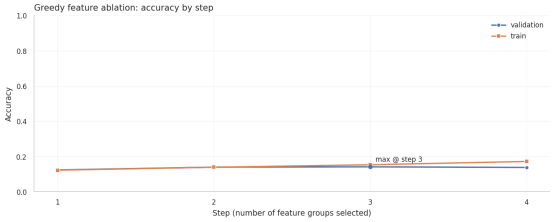
[2 pt] Explain why count is informative and *how* the $[0, 1]$ normalisation is implemented and used.

Feature Exploration. We repeat the greedy subset selection used for regression (*it is expected to reach this via trial-and-error from the students*), now maximising validation accuracy. As shown in Figure 5a, using $\tilde{c}$ alone reaches $\approx 13\%$ validation accuracy. Adding `atemp` (and its square) and coarse weather indicators nudges accuracy to ≈ 14–15% (on train, validation remains near ≈ 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 $\{\tilde{c}, \texttt{atemp}, \texttt{weathersit}\}$.

Points

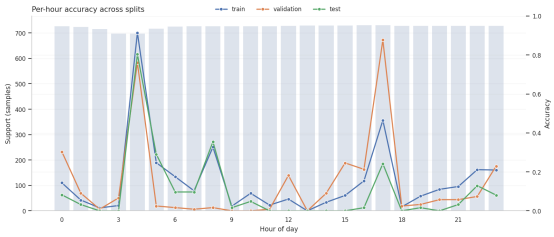
[2 pt] Show via trial-and-error (or automatic selection) which features are sensible to include and why.

Per-hour behaviour. Figure 5b 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.



Step	Chosen group
1	cnt_unit_interval
2	atemp, atemp^2
3	ws_mist, ws_light, ws_heavy
4	season_spring, season_summer, season_fall

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



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

Points

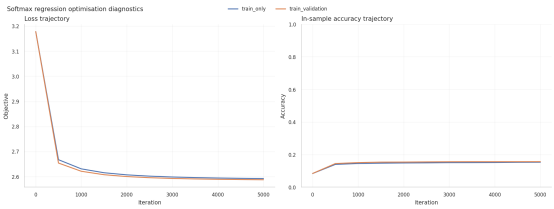
[2 pt] Provide split-wise accuracies with per-hour breakdown and discuss why certain hours are easier.

Optimisation diagnostics. Training/validation loss and in-sample accuracy flatten early (Figure 6a), 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.

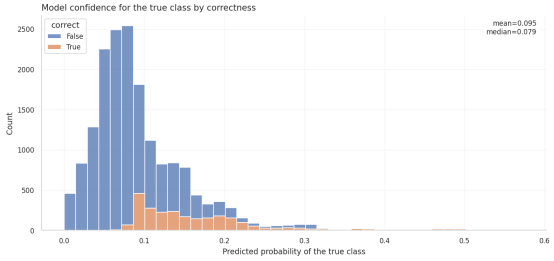
Points

[2 pt] Include loss & accuracy curves and interpret their plateau.

Confidence analysis. Figure 6b shows the predicted probability assigned to the *true* hour, stratified by correctness. Correct predictions have higher assigned probabilities than incorrect ones, but both 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.



(a) Softmax regression optimisation: objective and accuracy trajectories. Both curves plateau, more iterations do not really help.



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

Mutual information. From $P_\theta(y \mid x)$ one can estimate $I(Y;X) = \mathbb{E}x, \mathrm{KL}\big(P\theta(y \mid x), |, P_\theta(y)\big)$ by Monte Carlo over the evaluation split. Report the estimate and relate it to the improvement over uniform guessing $(1/24)$ as seen in Figure 7.

Points

[2 pt] Provide an MI estimate (or clear computation recipe) and relate it to the observed accuracy gain over chance.

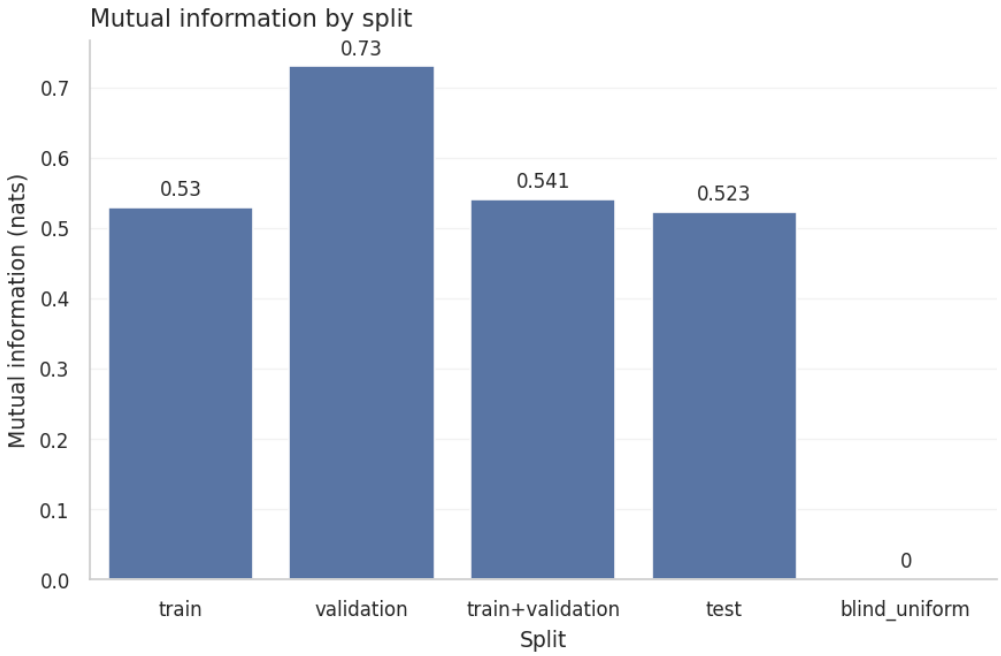


Figure 7: Showcase of the mutual information compared to the uniform guessing.

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.

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.