Practice 6: Regularization Analysis of NYC Taxi Data

Oleksandr Severhin

November 2025

1 Introduction and Data Preprocessing

This report analyzes the predictive drivers of taxi trip costs using the NYC Yellow Taxi dataset (January 2025). The objective was to predict the total trip cost (*log_total*) using a set of continuous and categorical predictors, comparing standard Ordinary Least Squares (OLS) against regularized regression methods: Ridge, Lasso, and Elastic Net.

The dataset underwent specific preprocessing steps to satisfy model assumptions and improve convergence:

- **Log-Transformation:** Due to right-skewness in monetary values and distances, we applied logarithmic transformations to '*total_amount*', '*trip_distance*', '*fare_amount*', '*tip_amount*' (+1), and '*tolls_amount*' (+1).
- **Encoding:** Categorical variables such as '*day_of_week*', '*payment_type*', and '*VendorID*' were One-Hot Encoded via the `model.matrix` function.
- **Splitting:** The data was partitioned into an 80% training set (n_{train}) and a 20% testing set (n_{test}) using `createDataPartition` to preserve distribution balance.
- **Standardization:** All predictors were automatically standardized (centered and scaled) by the `glmnet` package during the modeling phase to ensure fair penalization.

2 Model Performance Comparison

We fitted four model types: OLS (Baseline), Ridge ($\alpha = 0$), Lasso ($\alpha = 1$), and Elastic Net (optimized α). Hyperparameters (λ) were tuned using 5-fold cross-validation.

As seen in Table 1, the predictive performance across all models was remarkably similar. The OLS baseline achieved an R^2 of approximately 0.9717 and an RMSE of 0.0769 on the test set.

Table 1: Model Performance Metrics on Test Data

Model	**RMSE**	**MAE**	**R-Squared**	**Lambda**
OLS (Baseline)	0.0769	0.0577	0.9717	N/A
Ridge (λ_{min})	0.0769	0.0577	0.9717	≈ 0
Lasso (λ_{min})	0.0769	0.0577	0.9717	≈ 0
Elastic Net (λ_{min})	0.0769	0.0577	0.9717	≈ 0
Lasso (λ_{1se})	**0.0770**	**0.0577**	**0.9716**	**0.0010**

The cross-validation plots for Ridge and Lasso indicate that the Minimum MSE was achieved at extremely low values of λ (essentially converging to the OLS solution). However, by applying the "one-standard-error" rule (λ_{1se}), we selected simpler models with virtually no loss in accuracy. For instance, the Lasso (λ_{1se}) model saw a negligible drop in R^2 (from 0.97168 to 0.97162) while gaining parsimony.

The Elastic Net grid search revealed that the optimal mixing parameter was $\alpha = 1$, indicating that pure Lasso was preferred over a mix of Ridge and Lasso for this specific dataset.

The cross-validation plots for Ridge and Lasso indicate that the Minimum MSE was achieved at extremely low values of λ (essentially converging to the OLS solution). However, by applying the "one-standard-error" rule (λ_{1se}), we selected simpler models with virtually no loss in accuracy. For instance, the Lasso (λ_{1se}) model saw a negligible drop in R^2 (from 0.97168 to 0.97162) while gaining parsimony.

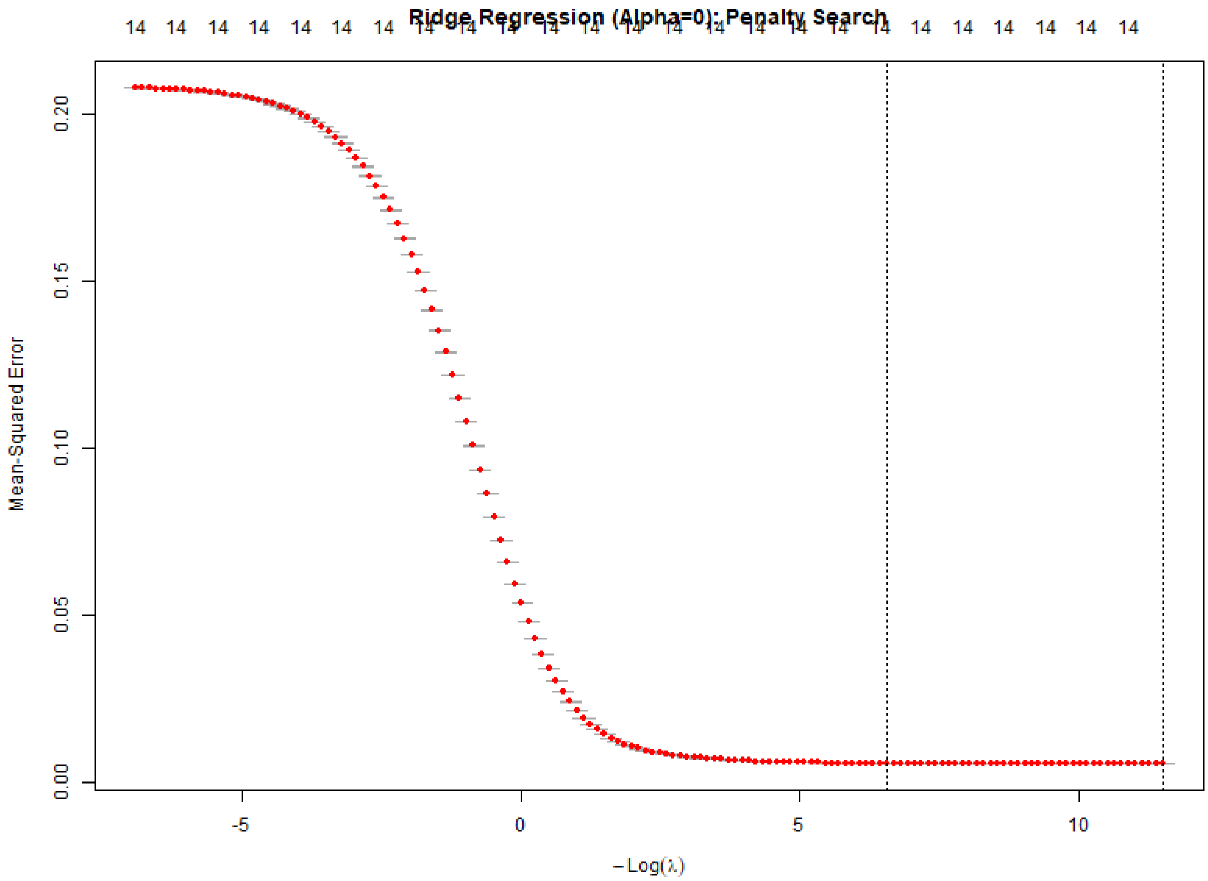

Figure 1: Ridge Regression Cross-Validation Plot. The red dots represent the Mean-Squared Error with error bars. The vertical dashed lines indicate λ_{min} and λ_{1se}.

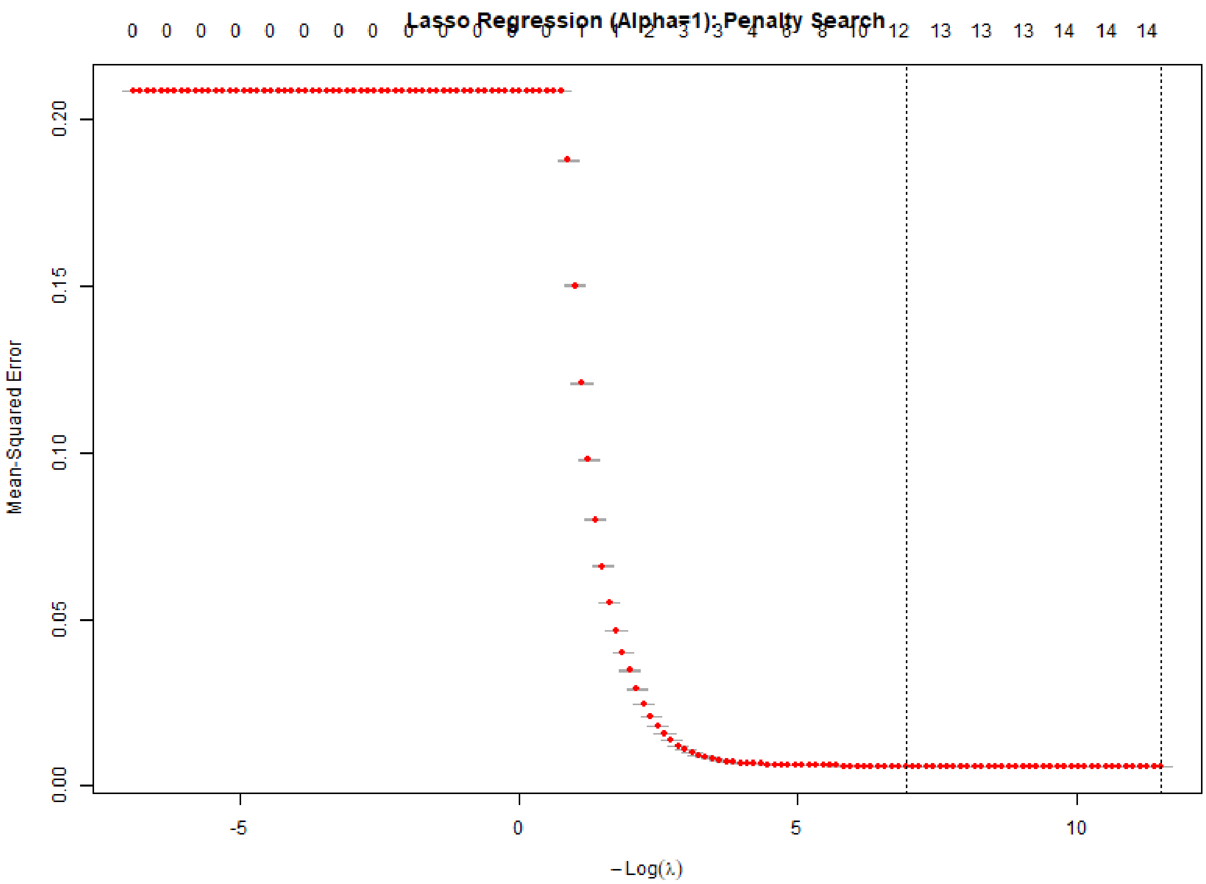

Figure 2: Lasso Regression Cross-Validation Plot. The numbers across the top indicate the number of non-zero coefficients in the model at that specific λ.

The Elastic Net grid search revealed that the optimal mixing parameter was $\alpha = 1$, indicating that pure Lasso was preferred over a mix of Ridge and Lasso for this specific dataset.

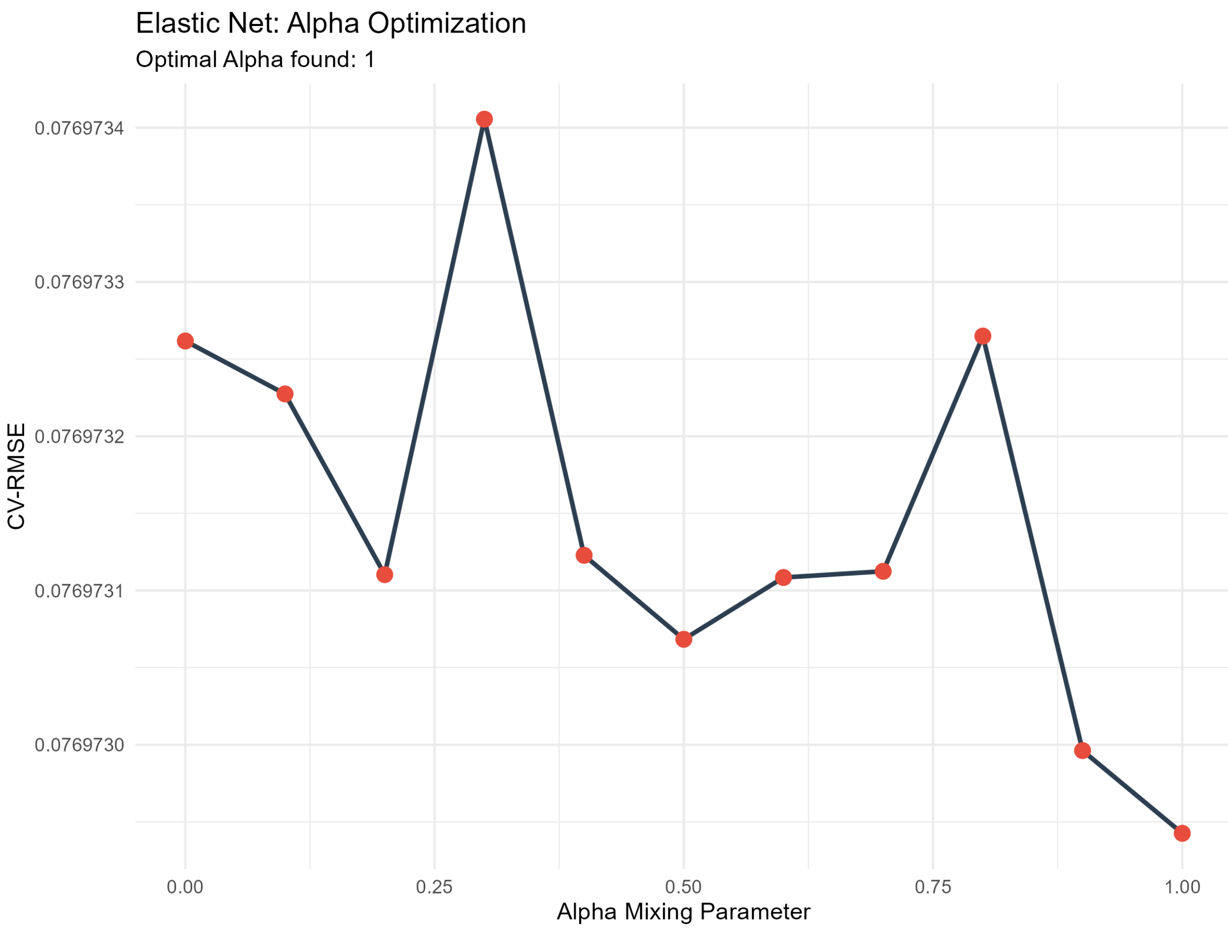

Figure 3: Elastic Net Alpha Optimization. The plot shows the cross-validated RMSE for different values of the mixing parameter α. The minimum is achieved at $\alpha = 1$ (pure Lasso).

3 Coefficient Shrinkage and Interpretation

The value of regularization in this analysis is best observed in variable selection rather than raw accuracy. Table 2 compares the coefficients of the full OLS model against the regularized versions.

Table 2: Coefficient Comparison (Selected Variables)

Term	OLS	Ridge (λ_{min})	Lasso (λ_{1se})	Elastic Net (λ_{1se})
(Intercept)	-0.0003	0.6721	0.6715	0.6711
log_distance	0.0172	0.0174	0.0146	0.0150
log_fare	**0.6447**	**0.6444**	**0.6492**	**0.6485**
log_tip	0.1437	0.1437	0.1386	0.1396
passenger_count	0.0015	0.0015	**0.0000**	0.0003
VendorID	-0.0002	-0.0002	**0.0000**	0.0000

Key Observations:

1. **Dominant Predictors:** ‘log_fare‘ is the most significant predictor (coef ≈ 0.65). This is intuitive, as the fare constitutes the bulk of the total cost.

2. **Variable Selection (Lasso):** The OLS model assigned small non-zero coefficients to '$passenger\_count$' (0.0015) and '$VendorID$' (-0.0002). The Lasso (λ_{1se}) model successfully identified these as noise, shrinking their coefficients to exactly zero. This simplifies the model: we can assert that, holding fare and distance constant, the number of passengers does not impact the log total cost.
3. **Shrinkage (Ridge):** The Ridge model retained all variables, merely shrinking the coefficients slightly. In this context, where true sparsity exists (irrelevant variables), Ridge is less interpretable than Lasso.

4 Peer Feedback Question

Question: "My Grid Search for Elastic Net resulted in an optimal Alpha of 1.0 (pure Lasso) and a performance identical to the standard Lasso model. Given that Elastic Net is computationally more expensive to tune than Lasso, under what specific conditions would you justify the extra effort of running a full Elastic Net grid search instead of just defaulting to Lasso?"

5 Answer to Exam Question

Answer: One should justify the extra computational effort of Elastic Net when there is high multicollinearity among groups of predictors, or when the number of predictors (p) exceeds the number of observations (n). In these scenarios, Lasso tends to arbitrarily select one variable from a correlated group and ignore the others, whereas Elastic Net (due to the Ridge L_2 penalty component) creates a "grouping effect," retaining all correlated variables and averaging their coefficients. Since my dataset ($p < n$) showed low multicollinearity among the core features, Lasso was sufficient, and the Elastic Net optimization correctly converged to the Lasso solution ($\alpha = 1$).