Lab Assignment Report

Assignment

CS503: Machine Learning

Name	Entry No.
Anshika	2021CSB1069

Instructor: Dr. Santosh Vipparthi

Teaching Assistant: Nishchala Thakur

Indian Institute of Technology Ropar
Punjab, India
March 11, 2025

Contents

1 | Overview

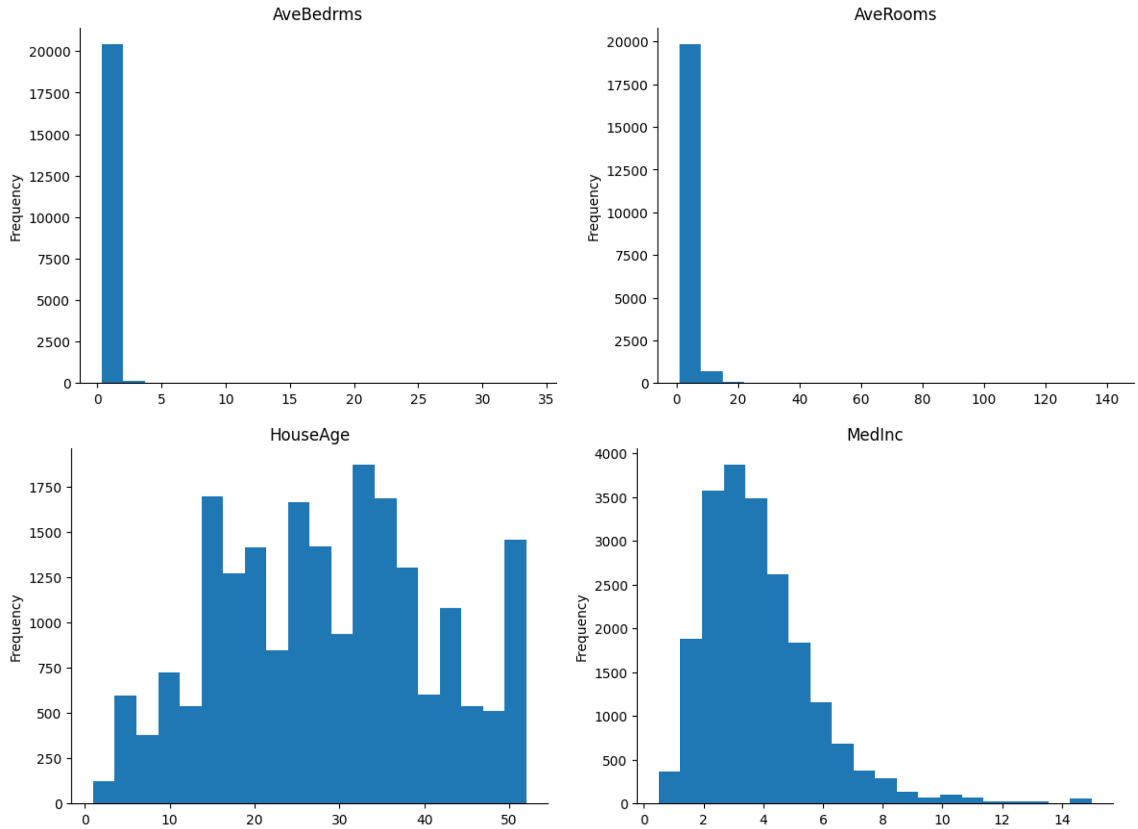
In this report an in-depth analysis of the application of L1 (Lasso) and L2 (Ridge) regularization techniques in linear regression is done. These regularizations are implemented and applied optimization algorithms (such as Gradient Descent) to train models on the California Housing Price dataset. A linear regression model with both L1 (Lasso) and L2 (Ridge) regularization to minimize overfitting.

2 | Dataset

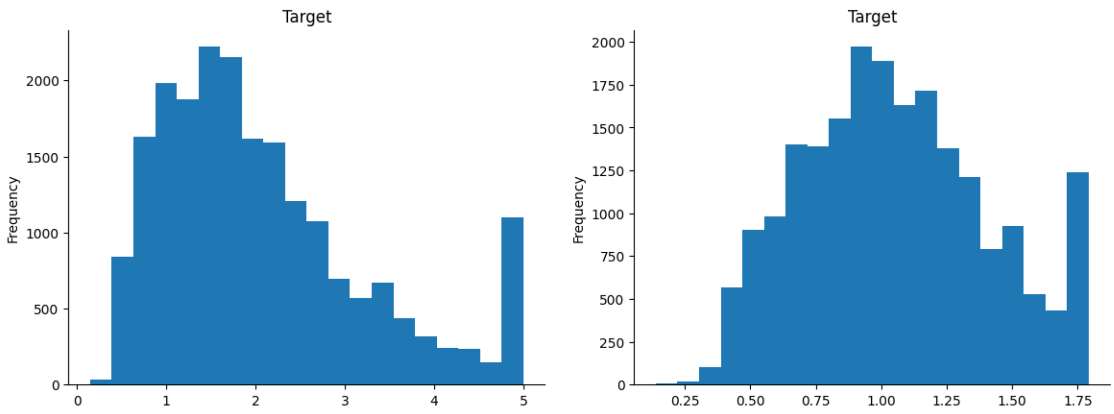
Following are the steps taken to ensure data quality and address any issues that could potentially affect model performance.

2.1 | Data Visualization

- **Histograms**: for independent variables

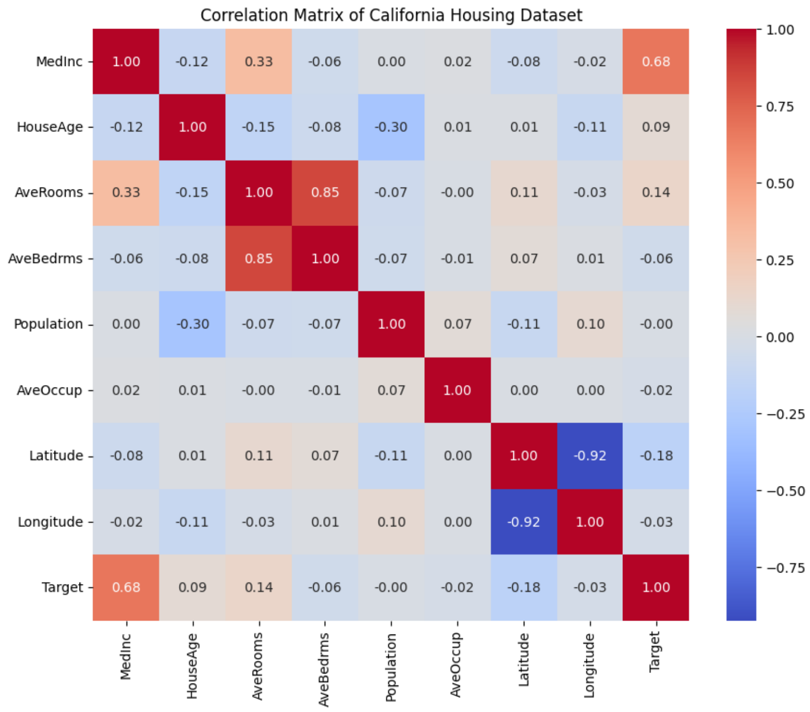


- **Distribution of Independent Variable:** target variable is skewed (left) so it is log transformed (right)



3 | Correlation Analysis

Correlation matrix was computed to identify any multicollinearity between the features. The resulting heatmap provided valuable insights into the strength and direction of relationships among features and the target variable ('Target').



3.1 | Model Evaluation

3.1.1 | Algorithm Considerations

- **Model Selection:** Linear Regression
- **Gradient Descent Regularization Variants:** To regularize the model, we implemented and compared two regularization techniques:
- **Lasso Regularization (L1)** L1 regularization adds the absolute value of the model coefficients to the cost function. It encourages sparsity, meaning that some coefficients are driven to zero, effectively performing feature selection. This is particularly useful when dealing with high-dimensional datasets.
- **Ridge Regression (L2)** L2 regularization adds the square of the model coefficients to the cost function, discouraging large coefficients but not setting them to zero. It reduces model complexity and helps prevent overfitting by shrinking the coefficients towards zero, but they remain non-zero.

- **Hyperparameter Tuning:**
  Various learning rates (`alpha`) and λ were tested for both regularizations. A grid search over the learning rates identified the best alpha for each method.

3.1.2 | Cost Function

The Cost function for both regularizations can be described as above:
L1 Regularization (Lasso) Cost Function:
$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} |\theta_j|$

L2 Regularization (Ridge) Cost Function:
$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2$
Where:

- y_i is the true label for the i-th sample.

This cost function was minimized across all training samples to obtain the optimal model parameters.

4 | Performance Measures

4.1 | Discussion on the effect of regularization on model performance and overfitting

1. **L1 Regularization (Lasso):** L1 regularization brings sparsity in the model, i.e. some feature weights may become zero, effectively removing less important features. This can be advantageous when you want a simpler model or when only a few features are important. However, it can lead to higher bias if it eliminates many important features.
2. **L2 Regularization (Ridge):** L2 regularization reduces the magnitude of the weights but never making them exactly zero. This helps to prevent overfitting by discouraging large coefficients values, with the model more generalizable. L2 is more stable than L1 for datasets where all features are important to contribute to the prediction.
3. **Effect on Overfitting**: Both L1 and L2 regularization add a penalty to the cost function to reduce the risk of overfitting. However, the choice between L1 and L2 regularization depends on the problem:
   - **L1 Regularization** can be particularly effective when you believe that only a subset of features are truly relevant.
   - **L2 Regularization** is useful when all features are useful but need to be shrunk to prevent overfitting.
4. **Model Performance:** In this specific case, Ridge (L2) regularization resulted in a slightly better average RMSE, showing that reducing the magnitudes of the coefficients,instead of forcing them to be zero, is more appropriate for the dataset .

4.2 | Visulaization of Regularization Techniques

- The cost function for both L1 4.1 and L2 4.2 regularization steadily decreases over time, indicating that the gradient descent optimization is working as expected.
- For **L1 regularization**, the cost curve initially decreases rapidly and then flattens, suggesting that it converges faster in the initial stages. This is typical for L1 due to the feature selection aspect, which can simplify the model early in training.
- For **L2 regularization**, the cost function decreases more gradually compared to L1. This behavior is expected, as L2 focuses on shrinking all coefficients, which may require more iterations to converge compared to L1.
- Both regularization techniques result in lower cost values as the epochs progress, indicating improved model performance with regularization. The eventual flattening of the curves suggests that the models have converged to their respective minima.

4.2.1 | Lasso Cost vs. iterations for different Learning rate and Regularization Coefficient with Lasso Regularization for 10000 epochs

Figure 4.1 demonstrates the effect of different learning rates and regularization coefficients on the Lasso Cost.

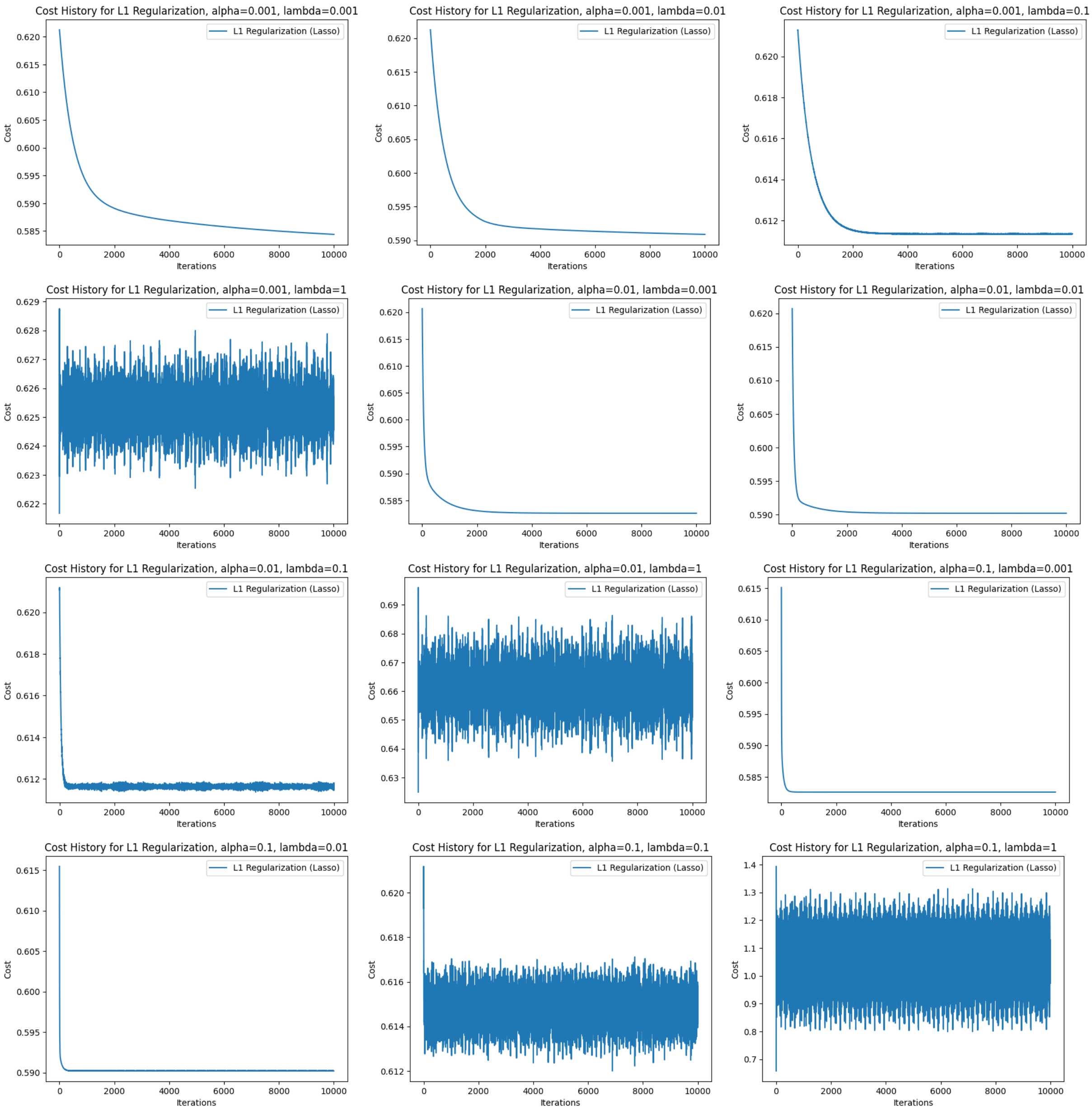


Figure 4.1: Lasso Cost vs Iterations

4.2.2 | Ridge Cost vs. iterations for different Learning rate and Regularization Coefficient for 10000 epochs

Figure 4.2 demonstrates the effect of different learning rates and regularization coefficients on the Ridge Cost.

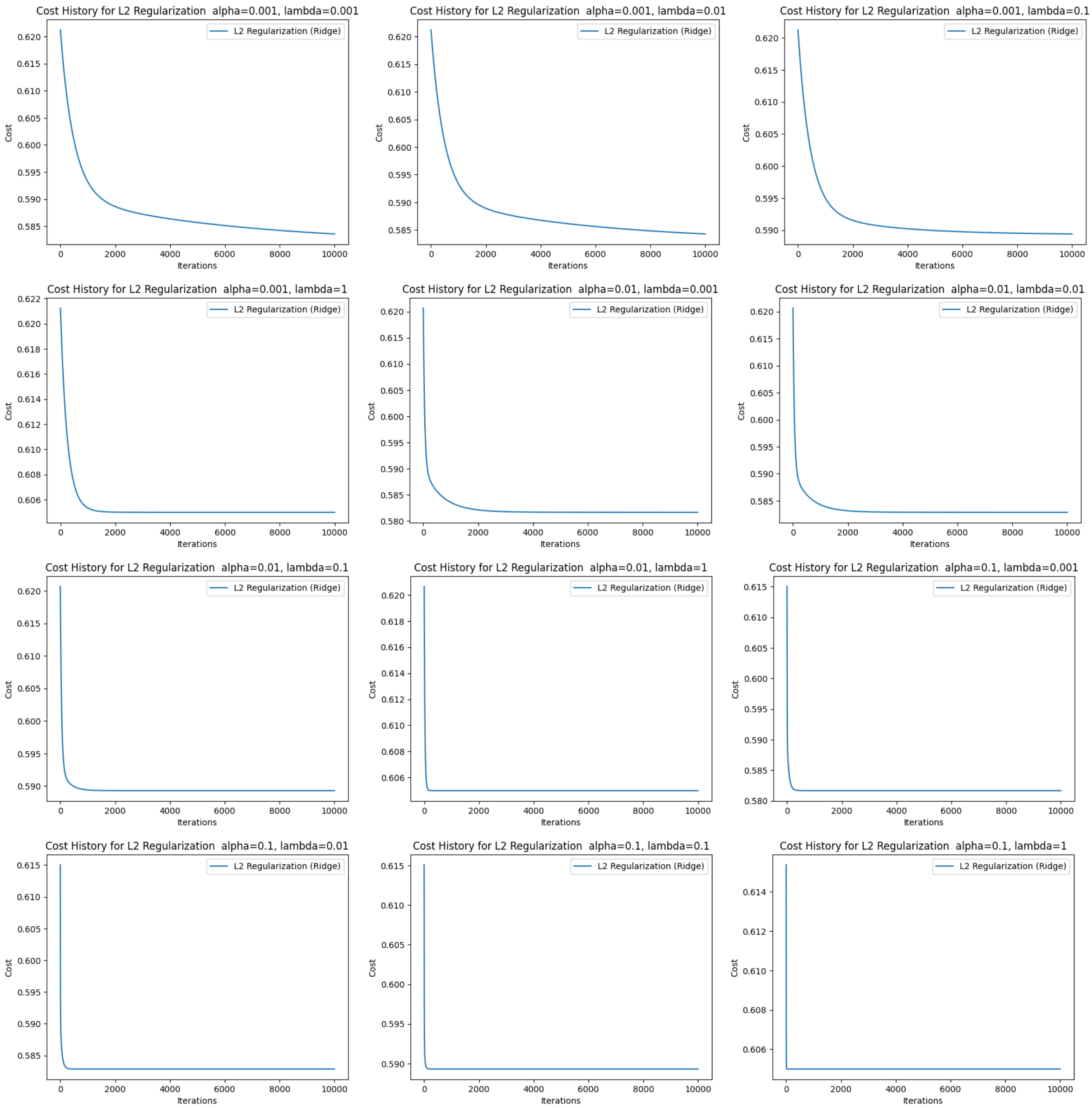


Figure 4.2: Ridge Cost vs Iterations

4.3 | Best Parameters after Grid Search and Results

The best RMSE values are given along with their respective costs in Table 4.1. Figure 4.3 represents the cost vs iterations for best α and λ.
The average RMSE values for each regularization type are:

- **L1 (Lasso)**: 2.2885
- **L2 (Ridge)**: 2.2568

Table 4.2 shows the RMSE for both L1 and L2 regularization types with different α andλ.

Regularization Type	Best RMSE	Best Alpha	Best Lambda
L1 Regularization (Lasso)	2.2284	0.1	0.001
L2 Regularization (Ridge)	2.2291	0.1	0.001

Table 4.1: Best RMSE, Alpha, and Lambda for L1 and L2 Regularization

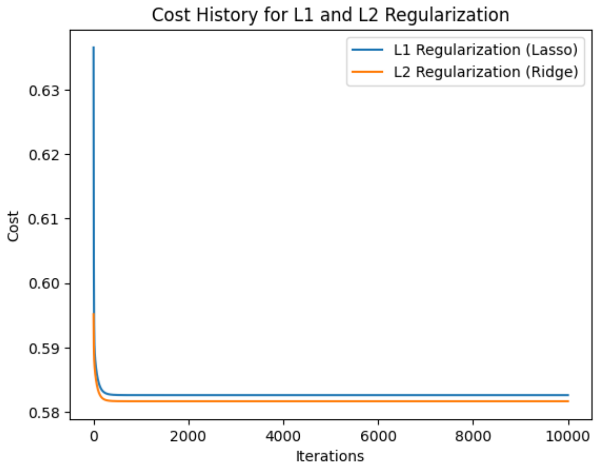


Figure 4.3: Cost History for L1 and L2 Regularization at best α and λ

Regularization Type	Alpha	Lambda	RMSE
L1 (Lasso)	0.001	0.001	2.2369911433
L1 (Lasso)	0.001	0.01	2.2474957820
L1 (Lasso)	0.001	0.1	2.3076448724
L1 (Lasso)	0.001	1.0	2.3631138313
L1 (Lasso)	0.01	0.001	2.2285131519
L1 (Lasso)	0.01	0.01	2.2422511867
L1 (Lasso)	0.01	0.1	2.3076317366
L1 (Lasso)	0.01	1.0	2.3627375606
L1 (Lasso)	0.1	0.001	2.2284289244
L1 (Lasso)	0.1	0.01	2.2422017516
L1 (Lasso)	0.1	0.1	2.3062932523
L1 (Lasso)	0.1	1.0	2.3887463471
L2 (Ridge)	0.001	0.001	2.2360810565
L2 (Ridge)	0.001	0.01	2.2378026765
L2 (Ridge)	0.001	0.1	2.2525945114
L2 (Ridge)	0.001	1.0	2.3109841452
L2 (Ridge)	0.01	0.001	2.2291637409
L2 (Ridge)	0.01	0.01	2.2308691338
L2 (Ridge)	0.01	0.1	2.2509632099
L2 (Ridge)	0.01	1.0	2.3109841151
L2 (Ridge)	0.1	0.001	2.2290545524
L2 (Ridge)	0.1	0.01	2.2308455299
L2 (Ridge)	0.1	0.1	2.2509632095
L2 (Ridge)	0.1	1.0	2.3109841151

Table 4.2: RMSE for different combinations of Alpha and Lambda for L1 (Lasso) and L2 (Ridge) Regularization

5 | Conclusion

In this lab assignment, L1 (Lasso) and L2 (Ridge) regularization are applied for linear regression on the California Housing dataset. It demonstrates the effect of these techniques in controlling overfitting for improving model generalization. A regularization term into the cost function for both L1 and L2 penalizes large coefficients, and simplifies the model to prevent it from fitting noise in the data.
Grid search to optimize hyperparameters is performed for learning rate and regularization strength. L1 regularization exhibited sparsity by driving certain coefficients to zero, performing feature selection.
L2 regularization shrunk the coefficients towards zero but did not eliminate any of them, resulting in a model that is less prone to overfitting while still retaining the predictive contributions of all features. This was evident in the slightly lower average RMSE observed with L2, showing that it was more suitable for this dataset where all features had relevance.
Overall, both L1 and L2 regularization improved the model's performance by preventing overfitting as can be seen from the cost curves, with L2 regularization achieving marginally better results. The trade-off between variance and bias was successfully managed through careful tuning of the regularization parameters.