Objective

The goal of this project is to develop a logistic regression classifier to predict whether a user will make a purchase based on various features of their shopping session. Key features include page interactions, browser type, traffic type, and weekend activity. The dataset used includes user session data, and performance is evaluated using different feature scaling methods (No Scaling, MinMax Scaling, Mean Normalization, Z-Score Normalization) along with regularization.

Model Performance with Feature Scaling

- No Scaling: Accuracy = 0.6986, Precision = 0.6362, Recall = 0.9242, F1 Score = 0.7536
- The model performed well with a high recall but lower precision, indicating more false positives.
- MinMax Scaling: Accuracy = 0.6803, Precision = 0.6312, Recall = 0.8637, F1 Score = 0.7294
- Minmax produced a slight drop in accuracy but a more balanced precision and recall.
- Mean Normalization: Accuracy = 0.6726, Precision = 0.6495, Recall = 0.7464, F1 Score = 0.6946
- This method resulted in the lowest accuracy but had the most balanced precision and recall.
- Z-Score Normalization: Accuracy = 0.7837, Precision = 0.8474, Recall = 0.6908, F1 Score = 0.7611
- Z-Score achieved the best accuracy and highest precision, though it had a lower recall.

Regularization vs. Non-Regularization

- Without Regularization: Accuracy = 0.6738, Precision = 0.6308, Recall = 0.8341, F1 Score = 0.7184
- Without regularization, the model showed moderate performance but risked overfitting.
- With Regularization (λ=100): Accuracy = 0.6803, Precision = 0.6312, Recall = 0.8637, F1 Score = 0.7294

Regularization improved recall and slightly boosted overall performance by reducing overfitting.

Graphical Representation

The cost function for both regularized and non-regularized models was plotted. The regularized model converged more smoothly, indicating better control over complexity and generalization. The unseen data set had a much more aggressive difference in Regulation vs non-regulation.

Conclusion

- Best Scaling Method: Z-Score Normalization yielded the highest accuracy (0.7837) and balanced precision and recall well.
- Impact of Regularization: Regularization improved recall and stabilized performance, though the gains were modest.
- Areas for Improvement: The model's precision could be improved by experimenting with different regularization parameters or by trying other algorithms like decision trees or ensemble methods.