A Comparison of Classification Models for the Advertisement Data Set

Ilana Greenberg

Introduction

This project seeks to compare classification methods for targeted advertising, by analyzing whether a user's decision to buy a product after viewing an advertisement can be predicted from their personal attributes like age, gender and salary. The results could be valuable for recommending a future model to advertisers.

Data

The dataset, sourced from Kaggle, contains 400 data points with 5 features each.

Selected Features: Age, Gender, and Estimated Salary of user

Selected Target: Whether or not a user purchased the product after viewing its advertisement

Comparison of Models: With vs Without Standard Scaler

Negative Recall: 99%
Positive Recall: 19%
F1 Score: 81%

Negative Recall: 91%
Positive Recall: 79%
F1 Score: 90%

Though the non-scaled regression was better at preventing unnecessary advertising (higher negative recall), using the scaler had a better result overall (higher f1 score)

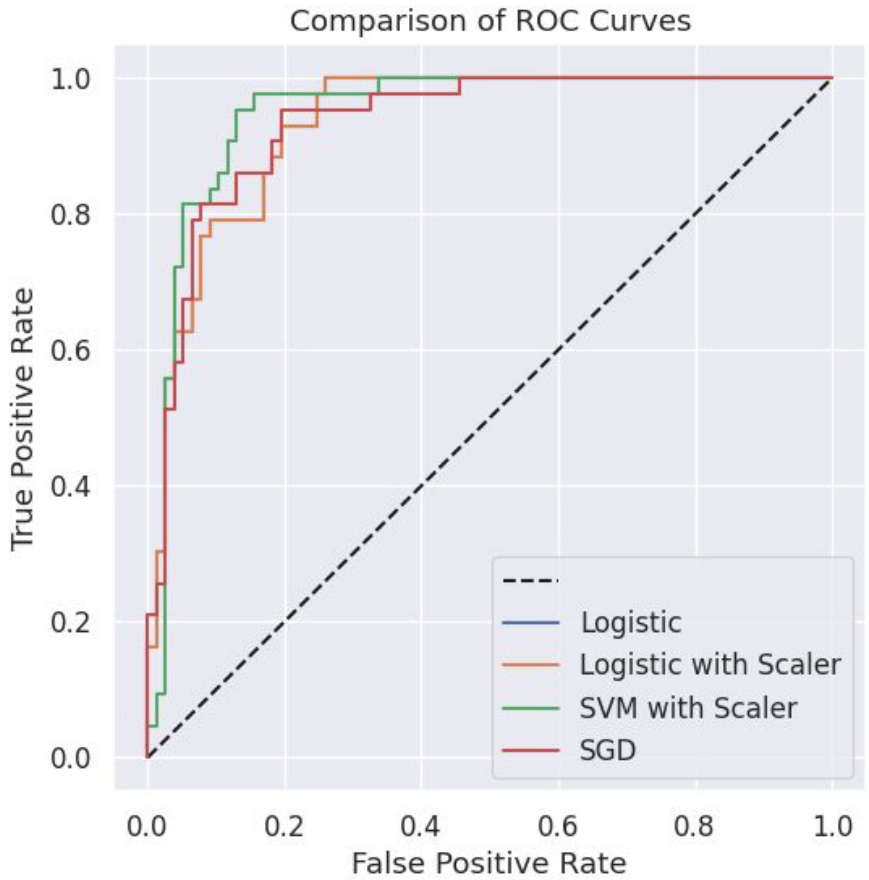
Comparison of Models: SVC and SGD

Negative Recall: 88%
Positive Recall: 91%
F1 Score: 91%

Negative Recall: 81%
Positive Recall: 91%
F1 Score: 87%

SVC was better at preventing unnecessary advertising (higher negative recall), and both captured the same percentage of true potential buyers (positive recall). So, SVC was slightly better overall (as evident by the F1 score)

Comparison of Models: ROC Curves



<u>Area Under Curve (AUC)</u>
Logistic: 73.8%
Logistic with Scaler: 93.4%
Support Vector Classifier (SVC): 94.5%
Stochastic Gradient Descent: 93.4%

The higher the AUC, the better a model is at maximizing true positive classifications and minimizing false positive classifications.

Logistic Regression performs better after a Standard Scaler is applied

SVC performs the best of all the models

Conclusion: Recommendations

The **Support Vector Classifier with a Standard Scaler** performed best.

Its AUC score (approx. 94%) was highest of all the models, showing that it was the best at maximizing the number of correct predictions while minimizing the number of incorrect predictions.

Confusion Matrix: The SVC was the best at minimizing the number of false negatives (its positive recall was 91%, compared to 19% for Logistic regression). However, it wasn't best at minimizing the number of false positives (its negative recall was 88%, while the Logistic Regression had a negative recall of 99%).

- If a company is concerned about only advertising to potential buyers, it might favor the Logistic algorithm with less false positives, though at the trade-off of loss of potential customers.
- The SVC algorithm captured the most potential buyers, but could lead to losing money over unnecessary advertising.
- However, SVC still has the best "balance" overall, as evident by the less stark difference in percentages