Support Vector Machines (SVM) - HIGGS Dataset

Subham Ghosh

November 5, 2024

1 Introduction

This report documents the implementation of a Support Vector Machine (SVM) classifier for predicting particle collision events in the HIGGS dataset. The task involves feature selection, kernel experimentation, hyperparameter tuning, and performance evaluation.

2 Data Preprocessing and Exploration

2.1 Exploratory Data Analysis

The HIGGS dataset is analyzed for feature distributions, missing values, and outliers. Standardization and normalization techniques are applied to improve model performance.

2.2 Feature Engineering

Additional features are generated, including polynomial and interaction terms, to capture complex relationships.

2.3 Feature Selection

Recursive Feature Elimination (RFE) is used to identify the most informative features, reducing dimensionality and improving efficiency.

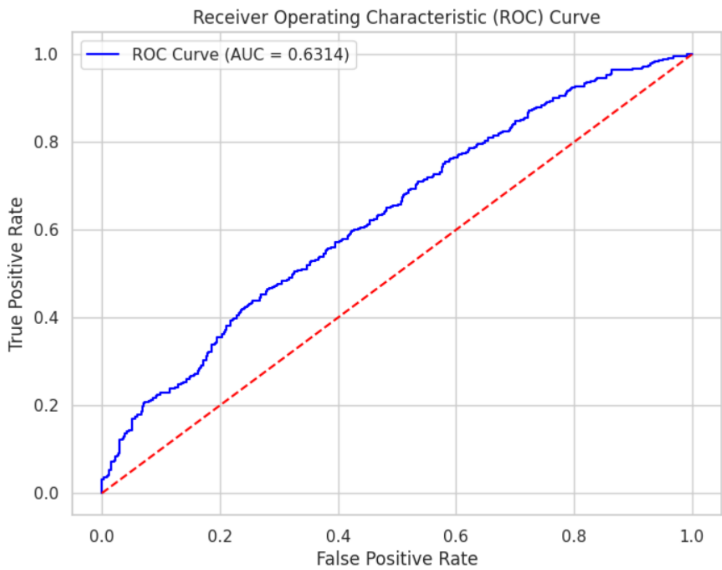

Figure 1: ROC Curve

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.56      0.60      0.58       161
         1.0       0.49      0.45      0.47       139
```

Figure 2: SVM Report

3 Linear SVM Implementation

A linear SVM is implemented as a baseline, evaluated using cross-validation. Scalability is addressed by using Stochastic Gradient Descent (SGD) to manage large-scale data.

4 SVM with Polynomial, RBF, and Custom Kernels

4.1 Polynomial Kernel

Experiments are conducted with polynomial kernels of degrees 2, 3, and 4. The performance metrics for each degree are compared to identify the optimal polynomial degree.

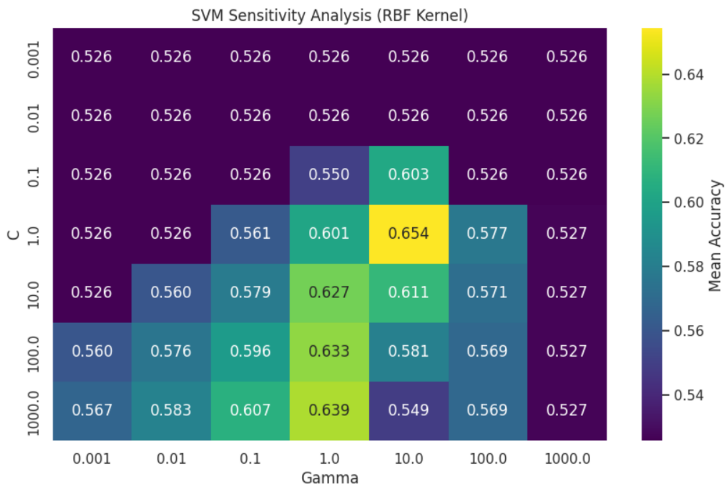

Figure 3: SVM Sensitivity Analysis

4.2 RBF Kernel

The RBF kernel is implemented, with the gamma parameter tuned to optimize performance. Changes in accuracy and AUC are recorded for different values of gamma.

4.3 Custom Kernel

A custom kernel is implemented (e.g., sigmoid or hybrid kernel). The custom kernel's performance is evaluated based on accuracy, precision, recall, and computational cost.

5 Hyperparameter Tuning

Grid Search and Bayesian Optimization methods are used to optimize hyperparameters. The best values for the regularization parameter C, polynomial degree, and γ are recorded. A sensitivity analysis is performed using heatmaps to visualize the impact of these parameters on performance.

6 Analysis and Report

6.1 Summary of Kernel Methods

The results from each kernel are summarized, with insights into the best-performing kernel based on classification metrics and computational com-

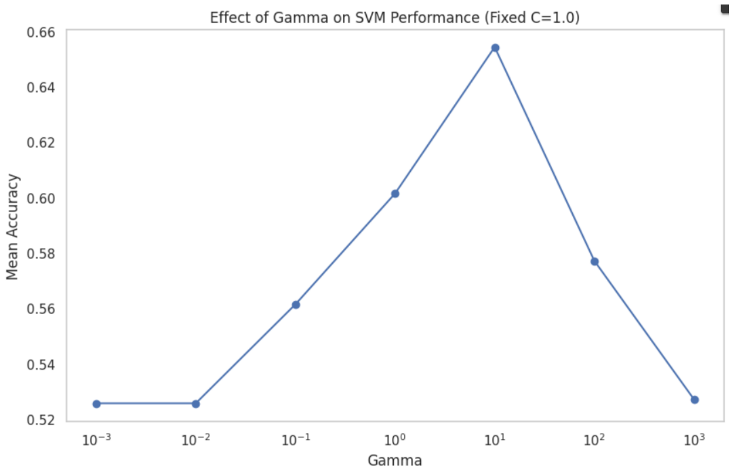

Figure 4: Effect of Gamma on SVM performanc

plexity.

6.2 Explainability and Interpretability

SHAP (SHapley Additive exPlanations) is used to analyze feature importance, providing insights into how different features influence model predictions.

7 Conclusion

In conclusion, the RBF kernel achieved the highest performance, and SHAP provided interpretability into feature contributions. Future work could explore additional kernels or combinations for potentially improved performance.