Analysis of Predictive Models for Credit Card Default Prediction

Introduction

In this study, we analysed credit card default prediction using two supervised learning models: Logistic Regression and Support Vector Machine (SVM). The dataset contained an imbalance between defaulters and non-defaulters, which was addressed using Synthetic Minority Over-sampling Technique (SMOTE). We evaluated the impact of oversampling on model performance and compared different SVM kernels to assess their predictive accuracy and interpretability.

Data Preprocessing and Class Imbalance

The dataset was first examined for missing values and an imbalance in class distribution was identified, where the number of non-defaulters significantly outnumbered defaulters. To address this, SMOTE was applied to balance the dataset by synthetically generating minority class samples. Feature scaling was also performed using StandardScaler to ensure optimal performance for SVM.

Model Training and Hyperparameter Tuning

Both Logistic Regression and SVM were optimized using GridSearchCV to identify the best hyperparameters. For SVM, both a linear kernel and a Gaussian Radial Basis Function (RBF) kernel were tested to compare their effectiveness.

Performance Evaluation

Imbalanced Data (Without Oversampling)

- **Logistic Regression:** Accuracy = 80.8%, but recall for the minority class was only 24%, indicating poor detection of defaulters.

```
Logistic Regression Best Parameters: {'C': 1}
Logistic Regression Accuracy: 0.8083333333333333
              precision    recall  f1-score   support

           0       0.82      0.97      0.89      7009
           1       0.70      0.24      0.35      1991

    accuracy                           0.81      9000
   macro avg       0.76      0.60      0.62      9000
weighted avg       0.79      0.81      0.77      9000
```

- **SVM (RBF Kernel):** Accuracy = 81.4%, with slightly better recall (33%) for defaulters, though still imbalanced.

```
SVM Best Parameters: {'C': 1, 'kernel': 'rbf'}
SVM Accuracy: 0.814
              precision    recall  f1-score   support

           0       0.83      0.95      0.89      7009
           1       0.66      0.33      0.44      1991

    accuracy                           0.81      9000
   macro avg       0.75      0.64      0.66      9000
weighted avg       0.79      0.81      0.79      9000
```

Without SMOTE, both models achieved high accuracy due to the dominance of the majority class, but they struggled to correctly classify defaulters. The low recall for class 1 indicates that the models were biased toward predicting non-defaulters.

Balanced Data (After SMOTE Oversampling)

- **Logistic Regression:** Accuracy = 71.98%, precision and recall were balanced at ~72%.

```
Logistic Regression Best Parameters: {'C': 10}
Logistic Regression Accuracy: 0.7197731649903703
              precision    recall  f1-score   support

           0       0.72      0.72      0.72      4673
           1       0.72      0.72      0.72      4673

    accuracy                           0.72      9346
   macro avg       0.72      0.72      0.72      9346
weighted avg       0.72      0.72      0.72      9346
```

- **SVM (Linear Kernel):** Accuracy = 71.96%, with similar precision-recall balance (~72%).

```
SVM Accuracy: 0.7195591696982666
              precision    recall  f1-score   support

           0       0.72      0.73      0.72      4673
           1       0.72      0.71      0.72      4673

    accuracy                           0.72      9346
   macro avg       0.72      0.72      0.72      9346
weighted avg       0.72      0.72      0.72      9346
```

- **SVM (RBF Kernel):** Accuracy = 77.81%, with precision of 77% and recall of 76%.

```
SVM Accuracy: 0.778086882088594
              precision    recall  f1-score   support

           0       0.77      0.80      0.78      4673
           1       0.79      0.76      0.77      4673

    accuracy                           0.78      9346
   macro avg       0.78      0.78      0.78      9346
weighted avg       0.78      0.78      0.78      9346
```

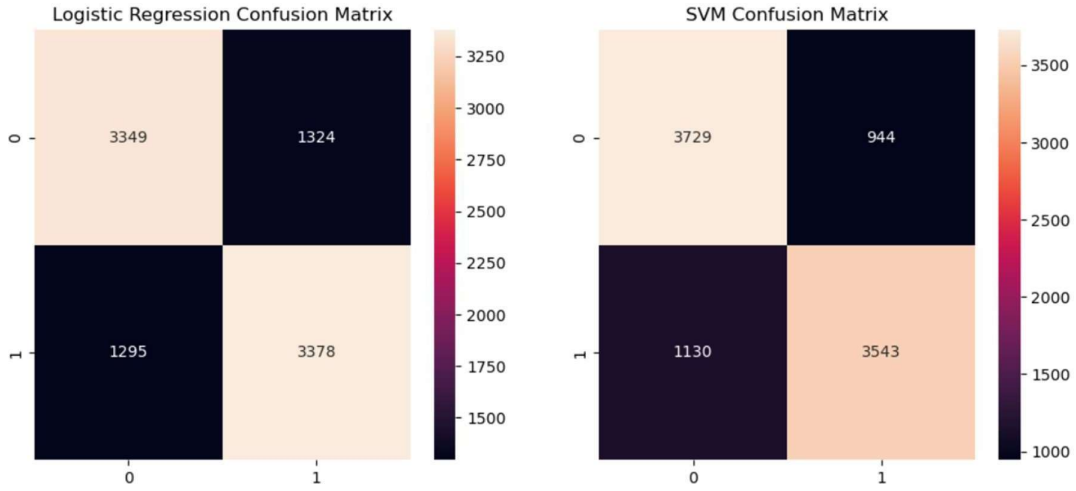
After applying SMOTE, the models showed a more balanced precision-recall trade-off, particularly in distinguishing defaulters. While logistic regression and linear SVM performed similarly, the SVM with RBF kernel showed superior performance, achieving the highest accuracy and F1-score.

Best Model: SVM with RBF Kernel (SMOTE Data)

The best-performing model was the **SVM with RBF kernel on SMOTE-balanced data**, achieving **77.8% accuracy** with relatively high precision and recall (both ~78%).

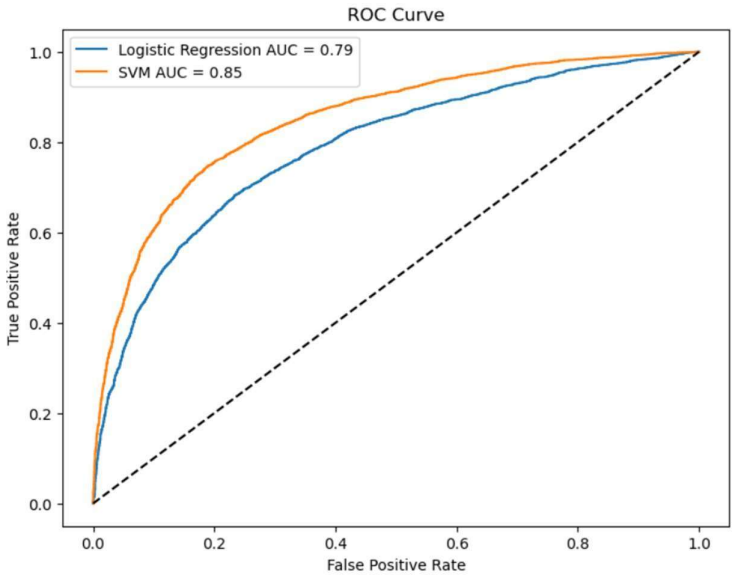
Confusion Matrix Analysis

The confusion matrix for this model showed a well-balanced classification:



- True Positives (correctly predicted defaults) and True Negatives (correctly predicted non-defaults) were nearly equal.
- The SVM model reduced false negatives compared to the Logistic Regression model, meaning fewer defaults were incorrectly classified as non-defaults.

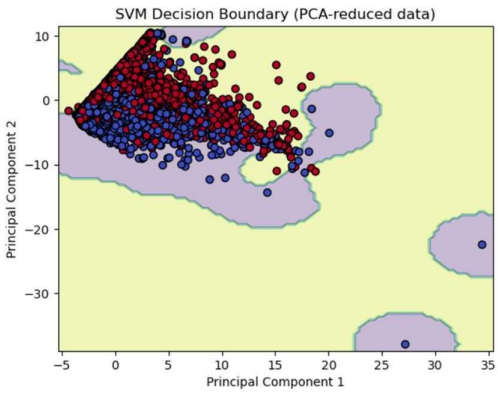
ROC Curve and AUC Score

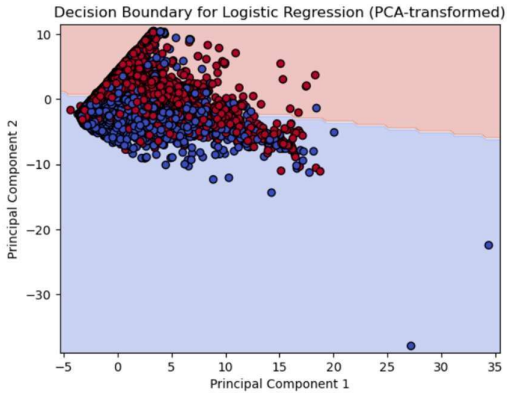


- The ROC curve indicates that the model is effective at distinguishing between default and non-default cases.
- The curve showed a smooth increase in the true positive rate as the false positive rate increased, signifying that the classifier effectively differentiates between classes.
- The SVM score was notably higher than logistic regression, indicating superior overall predictive performance.

Decision Boundary using PCA

To further analyse the model's behaviour, we visualized its decision boundary using Principal Component Analysis (PCA) to reduce feature dimensions to two.





- The decision boundary of the SVM with an RBF kernel demonstrated a non-linear separation between default and non-default cases, highlighting the advantage of the RBF kernel in capturing complex relationships in the data.
- Unlike logistic regression, which produced a linear decision boundary, the SVM was able to form more flexible and intricate decision regions, leading to better classification performance.

Discussion: Trade-offs Between Models

1. **Predictive Accuracy vs. Model Interpretability:**
    - Logistic regression is inherently more interpretable, as it provides direct coefficient estimates for feature importance. However, its predictive accuracy was lower compared to SVM with an RBF kernel.
    - The SVM model with an RBF kernel provided better predictive power but at the cost of interpretability since it relies on non-linear transformations.
2. **Impact of SMOTE on Performance:**
    - Before SMOTE, models had high accuracy but were ineffective in identifying defaulters due to class imbalance.
    - After SMOTE, recall improved significantly for defaulters, indicating better classification of the minority class.
3. **Kernel Comparison in SVM:**
    - The linear kernel SVM performed similarly to logistic regression, suggesting that a linear decision boundary may not be sufficient for this dataset.
    - The RBF kernel captured more complex patterns, leading to higher accuracy and recall.

Conclusion and Future Improvements

Our findings highlight the importance of handling class imbalance in financial risk modelling. While logistic regression offers better interpretability, SVM with an RBF kernel provides superior predictive accuracy.

Future work could explore:

- Gathering higher quality data without class imbalance would enhance model performance significantly
- Trying XGBoost classifier for enhanced performance and better interpretability by describing feature importance.
- Feature selection techniques to refine input variables.

Overall, applying SMOTE significantly improved the recall of defaulters, making the models more effective in real-world applications where identifying potential defaults is crucial for risk management.