Credit Cards Approval Classification Using SVM

Anna Abrahamyan [1] **Grisel Quispe Aramayo**[1] **Zahir Ahmad**[1]

Table 1. Contribution Table

	Part 1	Part 2
Zahir Ahmad	95%	30%
Grisel Quispe Aramayo	95%	30%
Anna Abrahamyan	95%	30%
Generative AI (model used)	5%	10%
Total	100%	100%

1. Introduction

Predicting credit card approval is a crucial task for financial institutions, which aim to balance risk management with customer acquisition. The adoption of machine learning in this domain holds the promise of automating decision-making processes, enhancing both the efficiency and the fairness of credit evaluations.

The dataset comprises a mix of numerical and categorical attributes, reflecting applicants' demographic and financial information. Through meticulous preprocessing, including one-hot encoding, handling of skewness, and feature normalization, we prepared the data for model ingestion, ensuring that each feature contributed meaningfully to the prediction task.

Support Vector Machines (SVM), a set of supervised learning methods well-suited for classification, were chosen as the primary algorithmic approach due to their capability in finding the optimal hyperplane that separates classes in a high-dimensional space. The study embarked on a comprehensive hyperparameter tuning for SVMs with various kernels—linear, RBF, and polynomial—seeking to unravel a configuration that excels in performance while maintaining generalizability.

2. Dataset Overview

The dataset comprises 690 entries, which includes the following types of variables:

- **Numerical Variables:** Age, Debt, YearsEmployed, CreditScore, ZipCode, and Income.
- **Categorical Variables:** Gender (binary), Married (binary), BankCustomer (binary), Industry, Ethnicity, PriorDefault (binary), Employed (binary), DriversLicense (binary), Citizen, and Approved (binary).

- **Demographics:** The dataset encompasses a wide spectrum of applicants with respect to **Age**, **Ethnicity**, and **Industry**, offering a comprehensive view of credit card application results.
- **Financial Information:** Variables such as **Debt**, **YearsEmployed**, **CreditScore**, and **Income** provide insights into the financial well-being and stability of the applicants.
- **Credit Approval:** The **Approved** column is binary and indicates whether an application was approved (1) or not (0), serving as the target variable for the modeling process.

The summary statistics underscore several notable points:

- **Age:** The age of applicants spans from 13.75 to 80.25 years, with a median age of 28.46 years.
- **Debt:** There is a substantial variation in debt levels among applicants, ranging from 0 to 28.
- **Employment and Credit:** The **YearsEmployed** and **CreditScore** variables imply a wide array of employment histories and credit behaviors, which are crucial in credit card approval decisions.
- **Income:** The income figures exhibit a broad range, from 0 to 100,000, with the median income considerably lower than the maximum, suggesting a distribution that is skewed.

3. Dataset Analysis

3.1. Class Distribution

The target variable **Approved** presents a slightly imbalanced distribution with 55.5% of the entries representing not approved applications (0) and 44.5% approved (1). This balance is crucial as it minimizes the likelihood of developing a model with a majority class bias.(See. Figure 1)

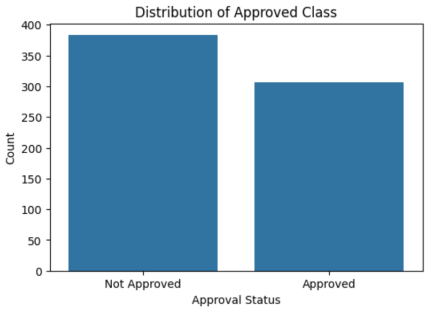

Figure 1. Distribution of Approved on Refined Dataset

3.2. Outliers and Distribution

Analysis of the box plots indicates the presence of outliers in several numerical features including **Age**, **Debt**, **YearsEmployed**, **CreditScore**, and **Income**. While outliers can often be indicative of data issues, in this context, they are considered to be representative of genuine variance in the applicant pool and thus will be retained for modeling.(See Figure 2)

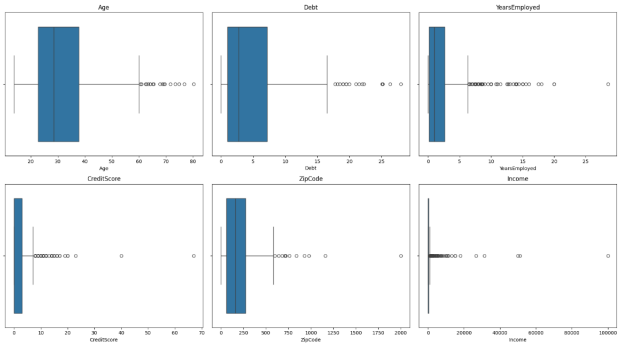

Figure 2. Outliers for numerical columns

3.3. Feature Correlation with Approval Status

The correlation analysis provides a rank of features based on their association with the approval status. The variable **PriorDefault** exhibits the highest positive correlation (0.72) with the **Approved** status, suggesting its significant impact on the outcome. This is closely followed by other influential features such as **Employed**, **CreditScore**, and **YearsEmployed**. Conversely, features like **ZipCode** and **Gender** demonstrate a weak correlation, implying limited predictive capabilities.

4. Data Prepossessing

Given the insights from the dataset analysis, our data preprocessing steps are carefully tailored to prepare the dataset for effective model training.

4.1. Categorical Feature Encoding

We began by encoding categorical features. For **Industry**, **Ethnicity**, and **Citizen**, we applied one-hot encoding to transform these into binary columns. In contrast, features like **Gender**, **Married**, **BankCustomer**, **PriorDefault**, **Employed**, and **DriversLicense** were already binary or treated as such, so no further encoding was necessary.

4.2. Feature Selection

Feature selection was approached both manually and through automated techniques. Utilizing a Random Forest model, we determined the initial ranking of feature importance, focusing on the features that have a more pronounced correlation with the target variable, **Approved**. The outcome reinforced the significance of **PriorDefault**, **Employed**, **CreditScore**, and **YearsEmployed** as top predictors. On the other hand, **Gender** and **DriversLicense** showed minimal importance, which corroborates the correlation analysis suggesting their exclusion.

Table 2. Selected Features and Their Impact Factor

Feature	Impact Factor
PriorDefault	0.272
CreditScore	0.094
YearsEmployed	0.094
Income	0.093
Debt	0.090
Age	0.078
Employed	0.063

4.3. Handling Skewness

With significant skewness observed in features like **CreditScore**, **YearsEmployed**, **Income**, **Debt**, and **Age**, we opted for log transformations to normalize their distributions. This step is pivotal for models that assume normality of the input features. Post-transformation, a notable improvement in distribution shapes was observed, indicating success in mitigating skewness.

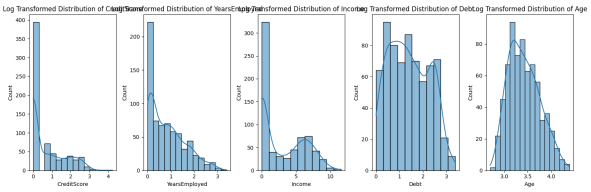

Figure 3. Skewness Handled

Table 3. Skewness Before and After Log Transformation

Feature	Skewness Before	Skewness After
CreditScore	5.152	1.059
YearsEmployed	2.891	0.819
Income	13.141	0.513
Debt	1.488	0.172
Age	1.167	0.421

4.4. Normalization

Finally, to ensure all features contribute equally to the model's prediction, we normalized the numerical features using the Min-Max scaler. This technique re-scaled the features into a range of [0, 1], thus preventing any single feature from disproportionately influencing the model due to scale.

4.5. Preprocessed Dataset

The dataset, now preprocessed, includes features engineered to reflect the true dynamics influencing credit card approval while also being formatted optimally for the modeling phase. Features such as **PriorDefault**, **CreditScore**, **YearsEmployed**, **Income**, **Debt**, and **Age** are expected to be significant predictors, with their distributions normalized to enhance the model's predictive power.

5. Model Training

5.1. Dataset Splitting

The dataset was divided into a training set and a testing set to evaluate the performance of the machine learning models. This split was performed with an 80-20 ratio. Stratification was applied during the split to maintain the proportion of the target variable classes in both training and testing sets.

5.2. Hyperparameter Tuning via Grid Search

To optimize the performance of the Support Vector Machine (SVM) model, a comprehensive grid search was conducted to find the best combination of hyperparameters. The parameters included in the search were the penalty parameter C with values {0.1, 1, 10, 100, 150, 200}, the kernel type with options {'linear', 'rbf', 'poly'}, and the kernel coefficient γ with values {0.001, 0.01, 0.1, 1, 10}. The grid search was executed using 5-fold cross-validation to ensure robustness of the results against overfitting and to assess the generalizability of the model across different data subsets.

The best parameter combination from the grid search yielded a C value of 0.1, a γ value of 10, and an 'rbf' kernel, achieving a cross-validation accuracy score of 0.862.

5.3. Linear SVM Model Training and Evaluation

For the Linear SVM model, grid search optimized the hyperparameter C at 0.1, balancing the model's ability to fit the training data while keeping a simple decision surface for better generalization. This configuration attained a cross-validation accuracy of about 0.837.

The model's classification performance was robust, as indicated by a classification report revealing a precision of 0.97 for the negative class and 0.84 for the positive class, denoting high accuracy in predictions. The recall for the positive class stood at an impressive 0.97, while for the negative class, it was slightly lower at 0.86, pointing to a disparity in identifying all non-approved cases accurately.

A PCA-based visualization of the decision boundary presented a clear but slightly overlapping separation of classes. (See. Figure 4)

Table 4. Linear SVM Classification Report and Best Parameters

Metric	Class 0	Class 1	Best Parameters
Precision	0.97	0.84	
Recall	0.86	0.97	C: 0.1
F1-Score	0.91	0.90	
Support	77	61	

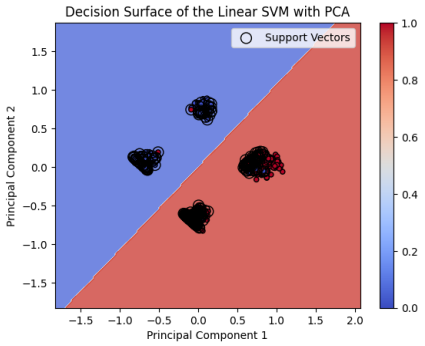

Figure 4. Decision Surface of the Linear SVM with PCA

5.4. RBF SVM Model Training and Evaluation

The grid search for the RBF SVM model yielded an optimal hyperparameter configuration with C set to 100 and γ at 0.1. A higher C value indicates a strong penalty on misclassifications, aiming for a lower bias at the potential cost of increased variance. The moderate γ value suggests the model's capacity to consider the influence of individual training samples while avoiding overfitting.

The model achieved an accuracy of approximately 0.857, a slight improvement over the Linear SVM. The precision-

recall balance shifted compared to the Linear SVM, with class 0 showing high precision (0.83) and class 1 demonstrating high recall (0.92). This implies that the RBF SVM is particularly adept at identifying approved cases but has room for improvement in correctly classifying non-approved cases, as evidenced by a lower recall for class 0 (0.75). (See. Figure 5)

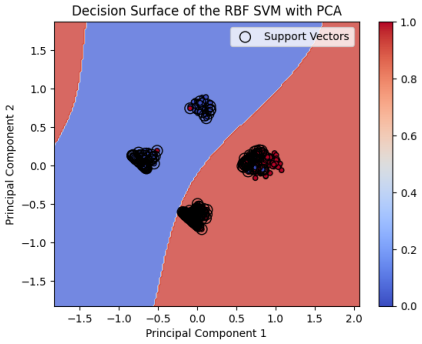

Figure 5. Decision Surface of the RBF SVM with PCA

Table 5. RBF SVM Classification Report and Best Parameters

Metric	Class 0	Class 1	Best Parameters
Precision	0.83	0.92	C: 100, γ: 0.1
Recall	0.95	0.75	
F1-Score	0.88	0.83	
Support	77	61	

5.5. Polynomial SVM Model Training and Evaluation

The Polynomial SVM, with hyperparameters set to $C = 100$, coef0 $= 0$, and degree $= 2$, achieved a commendable accuracy of approximately 0.864. This model, mirroring the RBF SVM in terms of precision and recall, presented a balanced trade-off with high recall for the negative class and high precision for the positive class. The classification report indicates a similar pattern as observed in the RBF SVM, reflecting the nuanced balance between the model's sensitivity and specificity.

The decision boundary visualized with PCA for the Polynomial SVM revealed an irregular separation of classes, less so than the RBF SVM, suggesting a more generalized approach while still accommodating some complexity. Despite this, there remains a risk of overfitting, as indicated by the non-linear and somewhat fragmented boundary. (See. Figure 6)

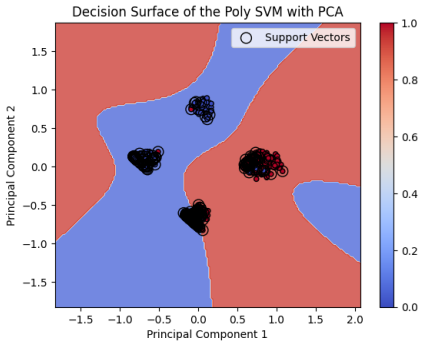

Figure 6. Decision Surface of the Poly SVM with PCA

Table 6. Polynomial SVM Classification Report and Best Parameters

Metric	Class 0	Class 1	Best Parameters
Precision	0.83	0.92	C: 100, coef0: 0, degree: 2
Recall	0.95	0.75	
F1-Score	0.88	0.83	
Support	77	61	

5.6. Model Performance Comparison

The radar graph elucidates that the Linear SVM (represented by the red line) achieves a balanced performance across all metrics, demonstrating its robustness and generalization capabilities.

The RBF and Polynomial SVMs, visualized with black and green lines respectively, demonstrate similar performances, suggesting that they both have comparable advantages and disadvantages. Their complex decision boundaries, while theoretically capable of modeling intricate patterns, do not necessarily lead to superior performance. In contrast, the Linear SVM, represented by the red line, shows a slight upper hand in the balance of metrics, reinforcing the findings from our earlier decision surface analysis. Essentially, the added complexity of the RBF and Polynomial SVMs does not translate into considerable gains, hinting that simpler models might be equally effective, if not more so, for this particular application.

5.7. Final Model Testing

The selected Linear SVM model was subjected to a final evaluation on the test dataset. The results from the classification report and ROC-AUC score are indicative of the model's robustness. It maintains high precision and recall across classes and achieves a ROC-AUC score of approximately 0.912, reflecting its strong discriminative power. Specifically, the precision for the negative class stands out at

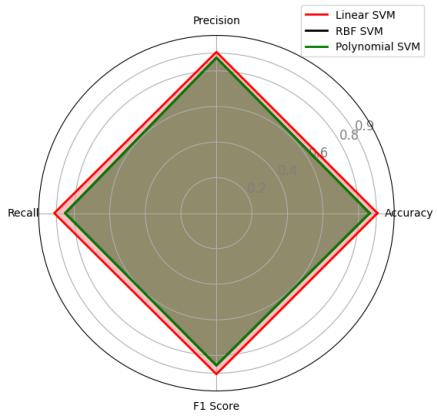

Figure 7. Radar Chart

0.97, coupled with an impressive recall for the positive class at 0.97. While the recall for the negative class is slightly lower at 0.86, it suggests potential areas for refinement.

Comparatively, the RBF and Polynomial SVM models experienced a decrease in accuracy on the test dataset, hovering around 85, which may point to overfitting during training and a lack of generalization to new data. This outcome aligns with observations from the decision surface plots, where the complexity of the decision boundaries could potentially be capturing noise rather than underlying patterns.

Table 7. Test Dataset Performance of the Final Linear SVM Model

Metric	Value
Precision (Negative Class)	0.97
Precision (Positive Class)	0.84
Recall (Negative Class)	0.86
Recall (Positive Class)	0.97
F1-Score (Negative Class)	0.91
F1-Score (Positive Class)	0.90
ROC-AUC Score	0.912

5.8. Critical Analysis and Future Directions

The performance of the RBF and Polynomial SVM models on the test dataset, with accuracies around 0.86 and similar precision-recall profiles, suggests a certain level of overfitting to the training data. While these models can delineate more complex decision boundaries, this does not necessarily lead to an improved performance on unseen data. In fact, both models showed a notable drop in recall for class 1 (approval), highlighting a potential weakness in identifying all positive instances.

The ROC-AUC scores for RBF and Polynomial SVMs further affirm the challenge of model complexity. Although the scores are commendable, the slight decrease compared to the Linear SVM model could be indicative of the cost of capturing noise along with patterns in the data.

For future work, the following steps could be explored to enhance model performance:

- Adjusting hyperparameters like γ for the RBF kernel and degree for the Polynomial kernel could further calibrate the models, especially in managing the trade-off between bias and variance.
- Incorporating cost-sensitive learning or different regularization techniques might benefit the models, particularly in handling the more complex patterns and decision boundaries observed.
- Since the recall for class 1 is notably lower in these models, techniques such as SMOTE for oversampling the minority class or targeted resampling could balance the recall metric across classes.
- Expanding the feature space or introducing new data preprocessing strategies could provide additional context for the models to learn from, potentially improving their ability to generalize.

6. Conclusions

This report has detailed the journey from initial dataset analysis through to the final model testing, providing insights into the credit card approval process using machine learning. During the model training phase, we employed Support Vector Machines (SVM) due to their effectiveness in binary classification problems. A grid search was meticulously conducted to fine-tune the hyperparameters for SVMs with different kernels: linear, radial basis function (RBF), and polynomial. The linear SVM emerged with a slight edge in performance, balancing accuracy and generalization effectively as evidenced by its classification report and ROC-AUC score on the test dataset.

The RBF and Polynomial SVMs, while capable of capturing more complex patterns, did not significantly outperform the Linear SVM when tested on unseen data, hinting at potential overfitting.

In conclusion, our findings underscore the effectiveness of SVMs in predicting credit card approvals, with a particular nod to the simplicity and robustness of the Linear SVM model.

The quest for an optimal balance between complexity and accuracy remains central to future iterations of this project. The insights garnered from this report lay a solid foundation for ongoing improvements and innovations in credit card approval predictions.