Assignment 1 - SVM Model Development using SAS Enterprise Miner on the Universal Bank Dataset

Tanushree Kumar | tanushree.kumar@outlook.com

DATA 640 - Fall 2024

Professor Steve Knode

University of Maryland Global Campus

Due: September 3, 2024

Introduction and Data Set Description

The objective of this analysis is to develop and evaluate six different Support Vector Machine (SVM) models using SAS Enterprise Miner to predict whether a customer will accept a personal loan based on the customer's demographic and financial data. The goal is to determine which SVM model provides the best performance in classifying whether a customer is likely to take the loan, addressing imbalances in the dataset while exploring model variations.

The problem domain revolves around personal loan marketing in the banking sector. Banks like Universal Bank provide various financial services and often promote personal loans to existing customers. However, not all customers may be inclined to accept such offers. By analyzing customer demographics, financial information, and historical data, predictive models can be created to identify potential loan acceptors. Such models are critical for banks to target marketing efforts efficiently, increase loan acceptance rates, and reduce costs associated with customer acquisition.

Support Vector Machine (SVM) models are well-suited for this analysis because of their ability to handle binary classification problems and complex, high-dimensional data. SVMs work by finding the optimal hyperplane that separates the data points into two classes (loan acceptors and non-acceptors). SVMs also allow for the use of kernel functions to capture non-linear relationships, which may be present in this dataset. This makes them ideal for distinguishing between customers who are likely to accept a loan offer and those who are not, even in the presence of overlapping or non-linearly separable data.

The dataset is sourced from the class portal and consists of 5000 rows and 14 columns. Each row represents a customer, and each column represents a demographic or financial attribute, except for the target variable. The target variable is Personal Loan, which is binary with

1 = customer accepted the loan, 0 = customer did not accept the loan. This variable is highly imbalanced, with a large proportion of customers declining the offer. The key features include:

- ID: Customer ID
- Age: Customer's age in completed years
- Experience: Number of years of professional experience
- Income: Annual income of the customer ($000)
- ZIPCode: Home address ZIP code
- Family: Family size of the customer
- CCAvg: Average spending on credit cards per month ($000)
- Education: Education level; 1 = Undergrad, 2 = Graduate, 3 = Advanced/Professional
- Mortgage: Value of mortgage if applicable ($000)
- Personal Loan: Target variable; Did this customer accept the personal loan offered in the last campaign?
- Securities Account: Does the customer have a securities account with the bank? (binary)
- CD Account: Does the customer have a Certificate of Deposit (CD) account with the bank? (binary)
- Online: Does the customer use internet banking facilities? (binary)
- Credit Card: Does the customer have a credit card issued by UniversalBank? (binary) (Knode, 2024a)

Data Cleansing and Preparation

Based on the dataset, approximately 10% of customers accepted the loan, while 90% did not. This imbalance in the target variable will be addressed in the modeling process using techniques such as resampling, cost-sensitive learning, or cutoff adjustment. The resampling

method uses oversampling or undersampling methods to balance the dataset. Cost-sensitive learning works by adjusting the cost function to penalize misclassification of the minority class. The cutoff adjustment works by altering the classification cutoff threshold from the default 0.5 value to better classify the minority class.

There were no missing values present in the dataset which meant no imputations were performed. There were also no outliers present which resulted in the variables having a skewness value between the [-2,2] range. Thus, no transformations were required either. The 'ZIPCode' column was removed as it wasn't deemed necessary to the models. Overall, minimal feature engineering was required since the dataset was in good condition.

Development of Predictive Models

The data was partitioned into a training and validation split of 70% and 30% respectively. Five initial SVM models were developed, each with different kernel functions and their respective default parameters. The first three models were made using the interior point optimization method and the last two models were made using the active set optimization method.

For all of these five models, a cost function with the value of 'c = 2' was also applied to adjust for the skewed target population. This was done using the Decision Table and the Decision Matrix. The cost function essentially forces the model to find and correctly classify the rare cases (Knode, 2016b). However, a caveat of this technique is that it may result in more false positives but will also identify more true positives. The value of 2 was chosen to emphasize the impact of class imbalance. The specific value of 2 was selected to show that misclassifying a loan acceptance, positive class, as a loan rejection, negative class, should be considered significantly more costly than the reverse error.

In many cases, misclassifying a minority class, such as predicting that a customer will not accept a loan when they actually would, might have more serious consequences. For example, if the bank wrongly rejects loan acceptance, it could result in lost business opportunities. Therefore, assigning a higher cost, in this case, 2, to false negatives ensures that the model becomes more sensitive to predicting the positive class accurately.

By assigning a higher cost to false negatives, the SVM will be penalized more for missing the minority class, effectively making the model more sensitive to it. This helps to address the imbalance by increasing the recall which is the ability to correctly predict the minority class.

Ideally, it would be good practice to experiment with different cost values and evaluate how they affect the model's performance. The goal is to find a balance that aligns with the business goals, maximizing the detection of the minority class without overly sacrificing accuracy in the majority class. Choosing the right cost value is essential as a false negative (rejecting a good loan) might have a higher financial impact than a false positive (approving a bad loan). Thus, the cost ratio should reflect the relative impact of these misclassifications.

Results

Model Number	**Model Name**	**Kernel Type**	**Kernel Value**	**Cost (C)**	**Accuracy**	**Sensitivity**	**ROC Index**
SVM 1	HP SVM (lin)	Linear	-	2	T = 0.9568 V = 0.9647	T = 0.6090 V = 0.6828	T = 0.96 V = 0.97
SVM 2	HP SVM (poly2)	Polynomial	2	2	T = 0.9877 V = 0.9807	T = 0.8985 V = 0.8690	T = 0.99 V = 0.99
SVM 3	HP SVM	Polynomial	3	2	T = 0.9951 V = 0.9760	T = 0.9582 V = 0.8690	T = 1.00 V = 0.98

	(poly3)						
SVM 4	HP SVM (RBF)	RBF	1	2	T = 0.9774 V = 0.9727	T = 0.7701 V = 0.7310	T = 0.99 V = 0.98
SVM 5	HP SVM (sig)	Sigmoid	[-1, 1]	2	T = 0.8488 V = 0.8329	T = 0.2000 V = 0.1172	T = 0.59 V = 0.54
SVM 6	HP SVM (poly4)	Polynomial	4	2	T = 0.9986 V = 0.9787	T = 0.9881 V = 0.9034	T = 1.00 V = 0.98
SVM 7	HP SVM (poly5)	Polynomial	5	2	T = 1.0000 V = 0.9700	T = 1.0000 V = 0.8759	T = 1.00 V = 0.98
SVM 8	HP SVM (RBF2)	RBF	2	2	T = 0.9723 V = 0.9727	T = 0.7164 V = 0.7241	T = 0.98 V = 0.98
SVM 9	HP SVM (RBF3)	RBF	3	2	T = 0.9588 V = 0.9627	T = 0.5701 V = 0.6207	T = 0.98 V = 0.98

Table 1. Table of model comparison with relevant aspects; (T = training; V = validation).

The models and their relevant aspects for model assessment can be seen above in Table 1. From the table, it can be observed that the polynomial and the RBF model had the best performance. This prompted four additional models that were developed using the polynomial and RBF kernels with updated kernel values. Two polynomial models were built with a kernel value of 4 and 5 respectively. The same was done for two more RBF SVM models with values of 2 and 3 respectively.

The table provides several key performance metrics for each SVM model: accuracy, sensitivity, and ROC index. These metrics evaluate how well each model performs, both in terms of overall correctness and its ability to handle imbalanced data. The accuracy measures the proportion of correct predictions (both true positives and true negatives) out of all predictions

made by the model. A high accuracy score indicates that the model is effective in predicting the overall outcome correctly, which can be seen for SVM 2 (poly2), which has the highest validation accuracy of 98.00%, indicating that this model performed the best in terms of predicting both classes correctly.

Sensitivity, also known as recall, measures the proportion of true positives correctly identified by the model. It is particularly important for imbalanced datasets, where the positive class (e.g., those who accepted the loan) may be underrepresented. High sensitivity means the model is effective at capturing positive cases. For example, SVM 7 (Poly5) has a sensitivity of 1.00 on the training set, indicating it captured all positives during training. However, in the validation set, its sensitivity drops to 0.8759.

The ROC index, or Area Under the Curve (AUC), measures the ability of the model to distinguish between the two classes (positive and negative). A higher value indicates that the model is better at differentiating between the classes. Models with an AUC close to 1 are considered excellent, while values closer to 0.5 suggest the model is no better than random guessing. Model SVM 7 (Poly5), with a ROC index of 1.00 on the training set and 0.98 on the validation set, performs exceptionally well.

Some key insights can be derived from the table. SVM 7 (Poly5) stands out as a top-performing model. It achieves a perfect accuracy (100%) and sensitivity (100%) on the training set and maintains strong performance on the validation set with an accuracy of 97.00%, sensitivity of 0.8759, and an excellent ROC index of 0.98. This suggests that the model generalizes well, striking a balance between capturing positive instances and avoiding overfitting. SVM 6 (Poly4) is also a strong contender, with a validation accuracy of 97.87%, a

sensitivity of 0.9034, and an ROC index of 0.98. This model performs slightly better in terms of sensitivity on the validation set than SVM 7, but overall, SVM 7 offers more consistent performance across all metrics. SVM 5 (Sigmoid) performs the worst across all models, with a validation accuracy of 83.29%, a sensitivity of just 0.1172, and an ROC index of 0.54. This model struggles particularly with sensitivity, indicating that it fails to capture the positive class effectively. Polynomial kernel models generally outperform the other kernels (linear, RBF, sigmoid) across most metrics. This suggests that the relationships in the dataset are best captured by higher-order polynomials, especially when handling non-linear patterns.

Conclusions

The objective of this analysis was to develop and evaluate multiple SVM models to predict customer acceptance of personal loans based on demographic and financial data. After testing a variety of SVM kernels, including polynomial, RBF, linear, and sigmoid, it was determined that polynomial kernels, particularly those with higher degrees, provided the best performance in terms of generalization and predictive accuracy. Based on the evaluation of accuracy, sensitivity, and ROC index, SVM 7 (Polynomial degree 5) emerged as the top-performing model. It consistently performs at a high level across training and validation datasets, demonstrating an excellent ability to generalize to new data while maintaining a good balance between sensitivity and overall classification accuracy. This model is particularly suitable given the nature of the data, which may contain complex non-linear relationships that a high-degree polynomial kernel is well-suited to capture. The significance of these metrics is critical, especially in dealing with imbalanced datasets. High accuracy ensures that most predictions are correct, but sensitivity is key to ensuring that the minority class (positive cases) is not neglected. The ROC index offers a robust measure of the model's ability to distinguish

between the classes, making SVM 7 a reliable and well-rounded model for this classification task.

Despite the strong performance of the polynomial SVM models, several limitations should be noted. The target variable in the dataset was highly imbalanced, with only 10% of customers accepting loans. While the use of cost-sensitive learning helped mitigate this issue, the dataset's imbalance might still influence the models' ability to generalize to real-world scenarios where the imbalance may be even greater. SVMs, particularly those using non-linear kernels, are considered "black-box" models, which limits the interpretability of their predictions. In highly regulated industries such as banking, where transparency is critical, this lack of interpretability could pose challenges. The analysis was performed on a specific dataset that may not fully capture all possible variations in customer behavior. As a result, the model might not generalize well to new, unseen data from different customer populations or external environments.

To address the data imbalance, future improvements could include resampling techniques to create synthetic samples of the minority class. This could further enhance the model's ability to detect loan acceptors. Implementing more interpretable models, such as decision trees or logistic regression with interaction terms, could provide business stakeholders with clearer insights into which features drive loan acceptance. While a fixed cost value of 2 was used, future iterations could experiment with a wider range of cost parameters and kernel configurations through grid search or Bayesian optimization to further refine model performance. Introducing additional features or external data, such as customer behavioral patterns or macroeconomic indicators, could provide the model with more contextual information, potentially improving its predictive power with the help of feature engineering.

References

HP SVM Node. (n.d.). Documentation.sas.com; SAS Help Center. https://documentation.sas.com/doc/en/emref/15.1/n18ip3imet0wokn1f39nqoxy9138.htm

Knode, S. (2016a). Adjusting for skewed target population [Vimeo]. In *Vimeo*. https://vimeo.com/186471846

Knode, S. (2016b). Cost function to adjust for skewed target population [Vimeo]. In *Vimeo*. https://vimeo.com/189827241

Knode, S. (2016c). SVM walkthrough [Vimeo]. In *Vimeo*. https://vimeo.com/189162413

Knode, S. (2017). Model Assessment - details [Vimeo]. In *Vimeo*. https://vimeo.com/225924138

Knode, S. (2020). SVM - demo with cutoff [Vimeo]. In *Vimeo*. https://vimeo.com/461049071

Knode, S. (2024a). *Universal bank description*. University of Maryland Global Campus. https://learn.umgc.edu/d2l/le/content/1226105/viewContent/33213198/View. Dataset description.

Knode, S. (2024b). *UniversalBank data*. University of Maryland Global Campus. https://learn.umgc.edu/d2l/le/content/1226105/viewContent/33213199/View. CSV file for the dataset.

Appendix A

All figures and visualizations mentioned in the report can be seen below.

Figure 1

Figure of the dataset.

Sample Table

ID	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Online	CreditCard	Personal Loan	Securities Account	CD Account
1	25	1	49	4	1.6	1	0	0	0	0	1	0
2	45	19	34	3	1.5	1	0	0	0	0	1	0
3	39	15	11	1	1	1	0	0	0	0	0	0
4	35	9	100	1	2.7	2	0	0	0	0	0	0
5	35	8	45	4	1	2	0	0	1	0	0	0
6	37	13	29	4	0.4	2	155	1	0	0	0	0
7	53	27	72	2	1.5	2	0	1	0	0	0	0
8	50	24	22	1	0.3	3	0	0	1	0	0	0
9	35	10	81	3	0.6	2	104	1	0	0	0	0
10	34	9	180	1	8.9	3	0	0	0	1	0	0

Figure 2

Figure of the dataset showing no missing values and summary statistics.

Name	ype	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
Age		.	0	23	67	45.3384	11.46317	-0.02934	-1.15307
CCAvg		.	0	0	10	1.937938	1.747659	1.598443	2.646706
CD_Account		2	0	.	.	.	.	.	.
CreditCard		2	0	.	.	.	.	.	.
Education		3	0	.	.	.	.	.	.
Experience		.	0	-3	43	20.1046	11.46795	-0.02632	-1.12152
Family		4	0	.	.	.	.	.	.
ID		.	0	1	5000	.	.	.	.
Income		.	0	8	224	73.7742	46.03373	0.841339	-0.04424
Mortgage		.	0	0	635	56.4988	101.7138	2.104002	4.756797
Online		2	0	.	.	.	.	.	.
Personal_Loan		2	0	.	.	.	.	.	.
Securities_Accou		2	0	.	.	.	.	.	.

Figure 3

Figure of the graph of the Chi-square values.

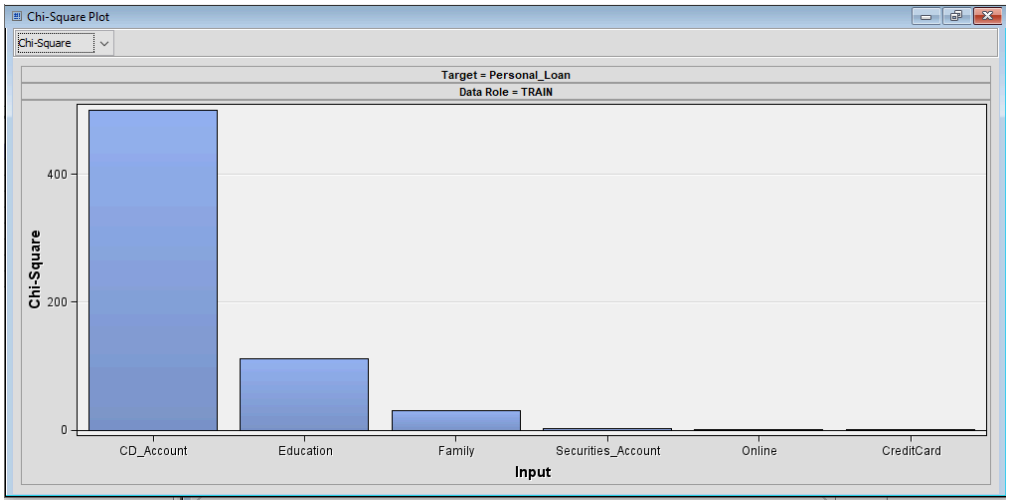

Figure 4

Figure of the graph of the variable worth.

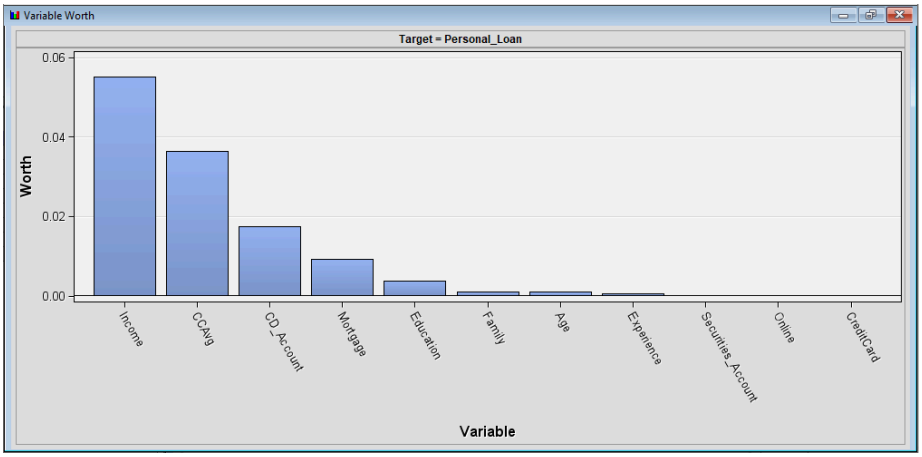

Figure 5

Figure of the cost matrix used for the cost function.

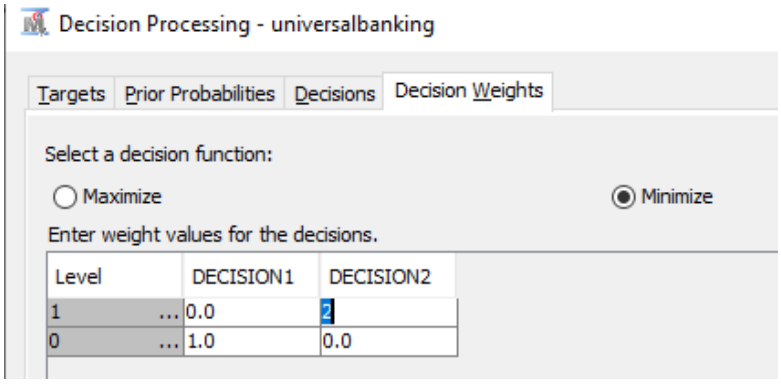

Figure 6

Figure of the ROC Chart for the first model comparison for the first five models.

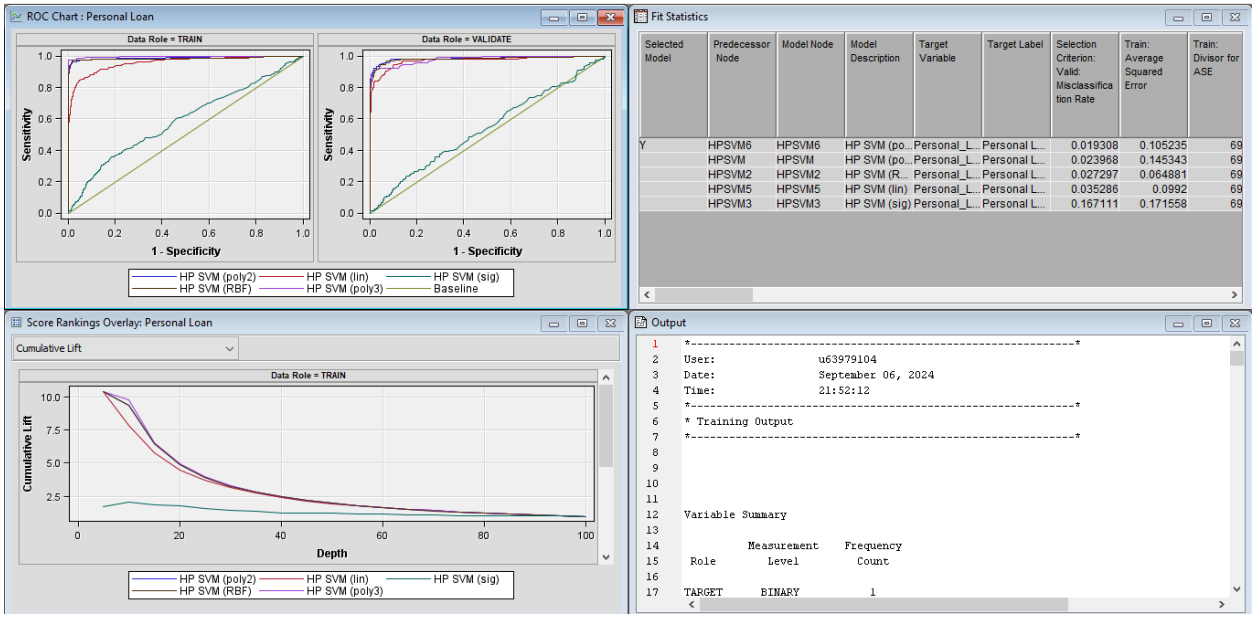

Figure 7

Figure of the ROC Chart for the first model comparison for the second four models.

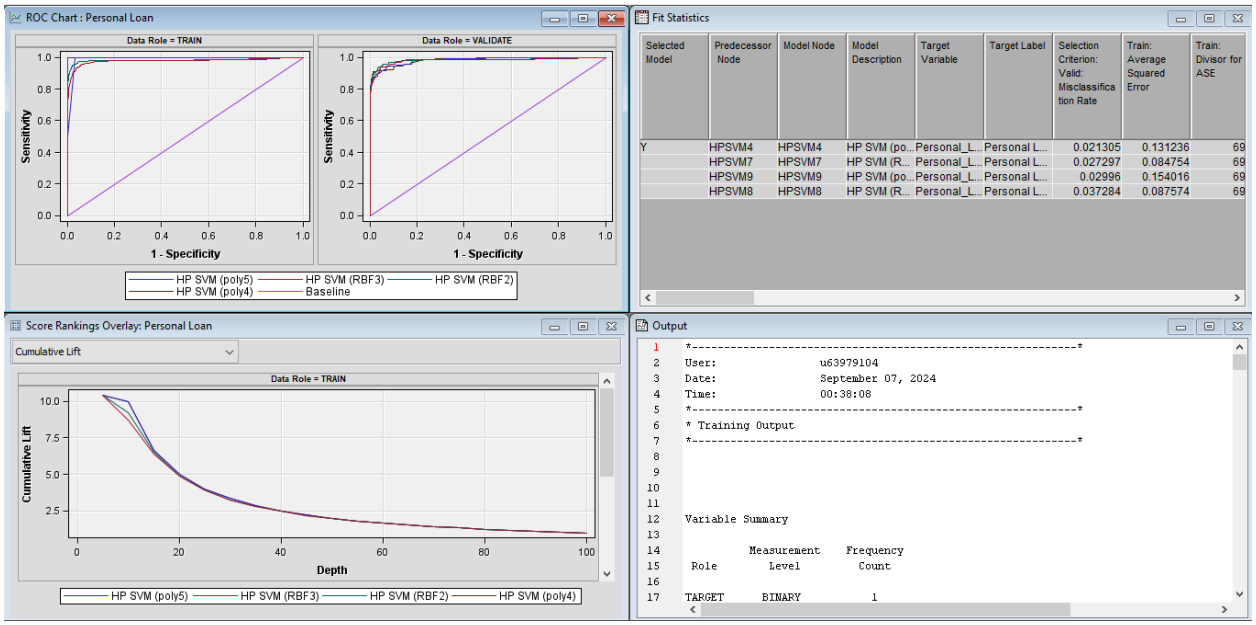

Figure 8

Figure of the cutoff curve for SVM 2.

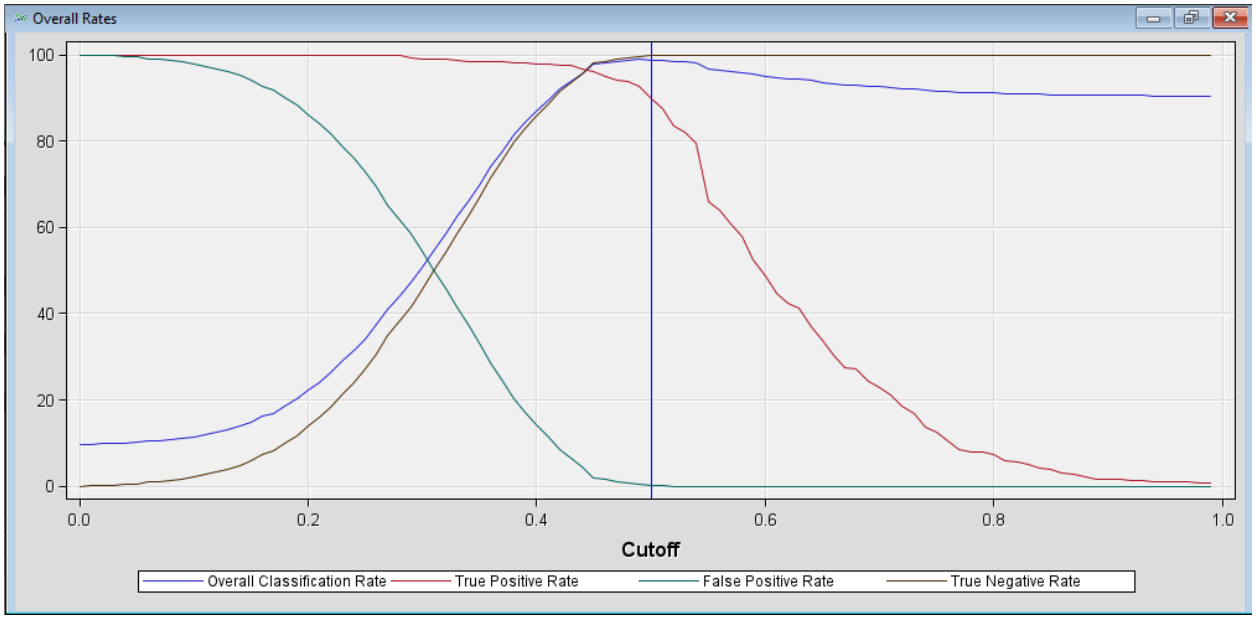

Figure 8

Figure of the adjusted cutoff curve for SVM 2 with value 0.45.

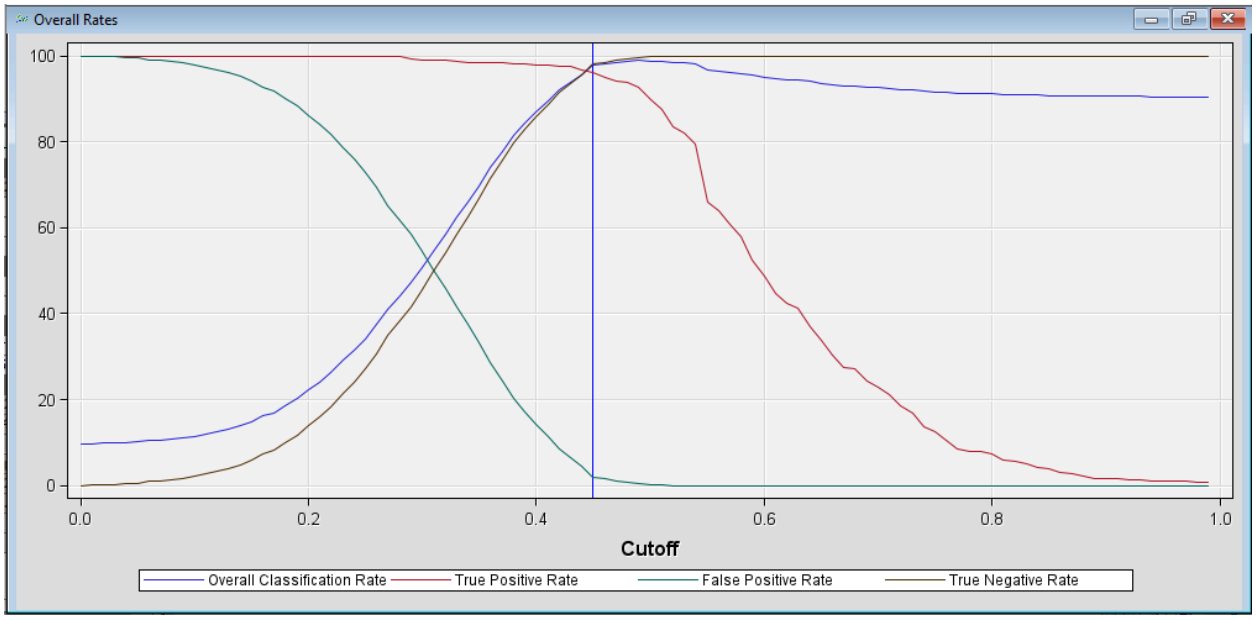

Figure 8

Figure of the cutoff curve for SVM 7.

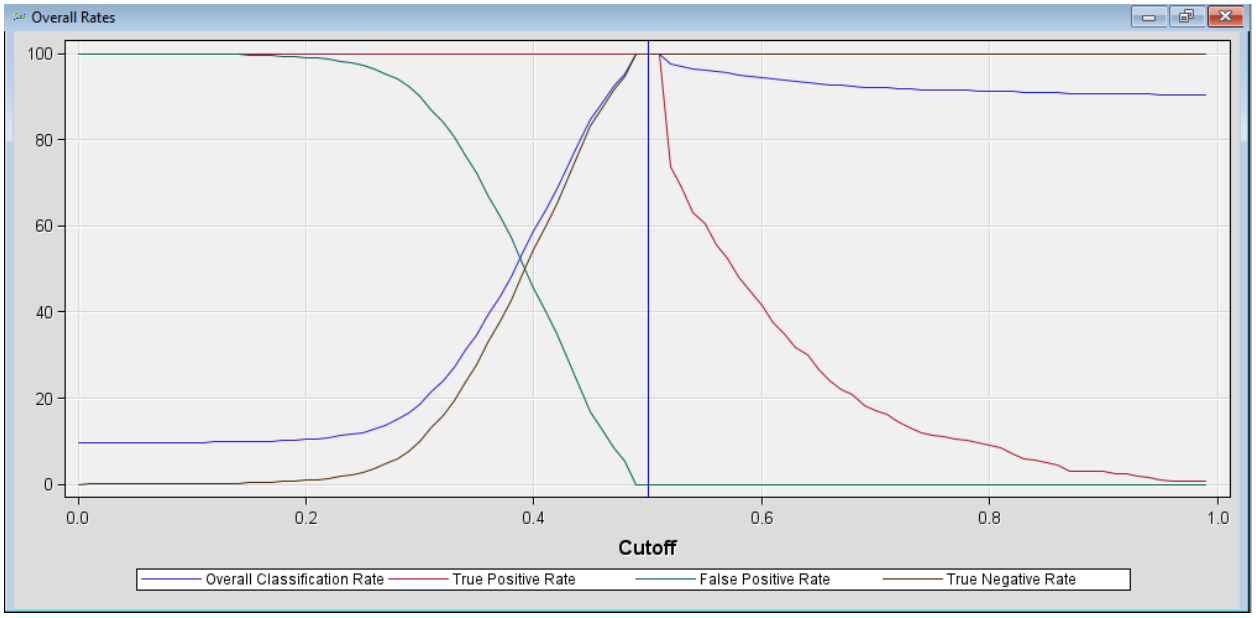

Figure 8

Figure of the SVM Model diagram.