Data 640: Predictive Modeling Fall 2023

Assignment # 5– SVM Models

Simon A Hochmuth

Email: shochmuth3@student.umgc.edu

Professor: Ed Herranz

Introduction

Churn is the percentage of accounts that cancel or choose not to renew a subscription, by doing so companies lose revenue due to the lost payments. For this reason, there is a vested interest for companies to correctly identify customers likely to leave and attempt to retain their subscription (ProductPlan 2021). The goal of this analysis is to utilize the data within the churn dataset to build and find the best predictive model that predicts whether a customer is likely to leave a company. For this analysis the Support Vector Machine (SVM) technique was used to build six predictive models within the SAS Enterprise Miner application. After the models are built, the six will be compared to see which had the best performance in predicting churn.

Data Set Description

As discussed, the dataset chosen had the goal of conducting predictive analysis to predict whether a customer is likely to leave the telephone company. The churn dataset contained 3333 observations and had 21 total variables, including the target variable of churn. There are no missing values in all the data. The original dataset provided by UMGC, shown in figure 1, has 21 variables consisting of customer personal data like phone number, plan type, state, and area code and individual plan metadata like number of voicemails, number of international calls, time of calls, number of extra charges for calls, and the total minutes. Overall, the variables had normalized data with the highest skew being 1.32 in the variable total international calls. For that reason, it was decided not to transform the data to make it more normalized. The target variable churn is a binary variable that states whether someone left or stayed with the telecommunication company. Where naturally the company cares to learn who is more likely to leave because that is

a lost revenue source. The target variable was highly skewed, with 14.5% of the data being TRUE or customers that left, and the other 85.5% being FALSE as shown in figure 2 below. Since the target variable is skewed, some work will need to be done to prepare it to be used by the model.

Data Cleansing and Preparation

As discussed earlier, the data in this model had normalized distributions with no missing data, meaning there was not a need to cleanse the data for the model. The first step was simply to pull the data into the SAS Enterprise Miner application, and to accurately set the roles and levels for each variable. Specifically, for churn the role was set to “target” and “binary” was set for the level. This is required for SVM models to work correctly. After the data was explored, the next step was to partition the data to prepare it for the models. It was decided to partition 30% of the data into the training data set, and 70% into the validation data set. After the data was partitioned, the next step was to move towards building the SVM models. It was decided not to transform the target variable at this point, because a cutoff node would be used to deal with the imbalance in the data after the models were created.

Predictive Models Developed

Table 1 below shows the 6 models that were developed for this analysis. Where the main deviation between each model was the optimization method used. Five of the models utilized an activation set optimization method, with one Radial Basis Function (RBF), two Sigmoid, and two Polynomials. The last model utilized a linear interior point optimization method. Each model parameter that was chosen can be seen in the table below.

As previously discussed, the target variable was skewed, with 14.5% of the 3333

observations being positive, and the other 85.5% being negative. For that reason, a cutoff node was implemented on each SVM model to address the imbalance in the target data. Since the optimization method and parameters differed between each model, the cutoffs were chosen on a case by case basis. The process of choosing the cutoff values included creating an initial node and setting the value equal to 14.5%, to match the current imbalance in the target data. It was decided that any optimal cutoff chosen would not go lower than this value, which was the case for only the linear model. This is to ensure all target true positives are represented within the data. To find the optimal cutoff value the charts, shown in figures 4 to 9 below, were analyzed with their supporting tables to provide the best cutoff value. The goal was to choose a cutoff that would maximize true positives without lowering model accuracy by a large amount. The final SAS diagram can be seen in figure 10 below and the optimal cutoffs chosen are listed in table 1 below.

Name	**RBF1**	**Poly2**	**Poly3**	**Linear**	**Sigmoid 1**	**Sigmoid 2**
Optimization Method	Active Set	Active Set	Active Set	Interior Point	Active Set	Active Set
Active Set Parameters	RBF = 1	2nd degree polynomial	3rd degree polynomial	N/A	Sig = +/- 1	Sig = +/- 2
Training Parameters	See Figure 3**	See Figure 3	See Figure 3	See Figure 3	See Figure 3	See Figure 3
Cutoff Value Chosen	0.2	0.37	0.44	0.14	0.27	0.14

*** Note: Training values were kept the same across all models*

Table 1: Shows the names of the 6 models used in SAS diagram, as well as the active set parameters, training parameters, and optimal cutoff values chosen.

Results

Now that the models were created, the last step was to analyze the results to compare which had the best performance in predicting churn. Since companies care more about customers that leave, a higher focus was put on true positives. Utilizing the SAS model comparison node, first the ROC curves for all models was compared, shown in figure 11 below. Models that perform better create 90 degree curves above a 45 degree baseline, while worse models follow the 45 degree or go below it. Following that logic, the ROC curve showed clearly that the 2nd and 3rd degree polynomials and the RBF model performed the best, while the two sigmoid and linear models performed poorly. Next the model comparison table was analyzed, shown in figure 12 and 13 below. First these confirmed that the RBF and two polynomial models did the best, however it showed that each model had between a 10% and 14% misclassification rate, or 90% to 86% accuracy, which is very close to the 14.5% to 85.5% skew in the target variable. This means there is a chance a model guessed that most values of churn = FALSE and would get very close to this accuracy. For that reason, the output was analyzed in figure 14 to understand the number of true positives and negatives. Figure 14 shows that the two Sigmoid, RBF, and Linear models had very low true positive predictions, meaning the misclassification rate of 14% was due to guessing negative for most values. When looking at the true positive predictions on the validation data set, these models were between 0% and 16% accuracy or 100% to 84% misclassification rate. Looking at the 2nd and 3rd degree polynomials, far more values were predicted to be positive with the 2nd degree having 46% accuracy, and the 3rd degree having 53% accuracy on the validation data set. Also, the 2nd degree model had a 3% misclassification rate for false positives and the 3rd degree model had an 8% misclassification rate. However as previously mentioned, companies would put higher focus on true positive accuracy, meaning the 3rd degree polynomial model with a 0.44 cutoff was decided to be the best for the purpose of this analysis.

As mentioned, companies naturally put higher focus on customers that are leaving the company because of the lost revenue streams that come from it. For that reason, a higher focus was put on accuracy in true positive values. That is why the 3rd degree polynomial came out on top, even with lower overall accuracy of 87% compared to the 90% overall accuracy on the 2nd degree polynomial.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Roc Index	Valid: Roc Index ▼	Valid: Misclassification Rate	Train: Misclassification Rate
	CUT3	HPSVM	RBF1_SVM	churn	0.939	0.88	0.140043	0.107214
Y	CUT2	HPSVM5	Poly2_SVM	churn	0.928	0.873	0.105353	0.07014
	CUT6	HPSVM4	Linear_SVM	churn	0.834	0.838	0.141328	0.133267
	CUT	HPSVM3	Poly3_SVM	churn	0.988	0.819	0.134904	0.017034
	CUT5	HPSVM2	Sig2_SVM	churn	0.59	0.594	0.145182	0.144289
	CUT4	HPSVM6	SIG1_SVM	churn	0.524	0.55	0.149465	0.153307

Figure 12: Shows statistical results of models, with the RBF and 2nd Degree Polynomial doing the best, and Sigmoid models doing the worst.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
HPSVM5	Poly2_SVM	TRAIN	churn		63	847	7	81
HPSVM5	Poly2_SVM	VALIDATE	churn		182	1932	64	157
HPSVM3	Poly3_SVM	TRAIN	churn		15	852	2	129
HPSVM3	Poly3_SVM	VALIDATE	churn		153	1834	162	186
HPSVM6	SIG1_SVM	TRAIN	churn		139	840	14	5
HPSVM6	SIG1_SVM	VALIDATE	churn		323	1970	26	16
HPSVM	RBF1_SVM	TRAIN	churn		105	852	2	39
HPSVM	RBF1_SVM	VALIDATE	churn		292	1961	35	47
HPSVM2	Sig2_SVM	TRAIN	churn		144	854	0	0
HPSVM2	Sig2_SVM	VALIDATE	churn		339	1996	0	0
HPSVM4	Linear_SVM	TRAIN	churn		128	849	5	16
HPSVM4	Linear_SVM	VALIDATE	churn		301	1967	29	38

Figure 14: Shows statistical results of models, with the RBF and 2nd Degree Polynomial doing the best, and Sigmoid models doing the worst.

Conclusions and Takeaways

The results of the 6 models showed that the 3rd degree polynomial with parameters described in table 1 above was the best model for the purposes of the analysis. The 3rd degree polynomial had 0.81 value for ROC, meaning it statistically was more significant than a baseline model. After more analysis, it was shown that it did better than the others in predicting the true positive values for churn. The 3rd degree polynomial model had 53% accuracy in correctly labeling the true positive churn values, and 92% accurate in correctly labeling true negatives. The runner up was the 2nd degree polynomial, but it was decided to be worse because it was only 46% accurate in predicting true positives. Other cutoff values were not assessed, because the highest performing value was individually chosen for each model. Therefore, the 3rd degree polynomial proved to add value to the telecommunication companies' analysis of customers leaving the company. For this reason, it is recommended for the company to utilize this model that predicts churn, with the caveat that there will need to be more work done to achieve greater performance in the model. Naturally, the company would want to have close to 100% accuracy in true positives which is why it is suggested to expand upon the model. While the model is being expanded, it is also a recommendation to start to target customers who live within the true positives to build retention. Even with 53% accuracy, there is close to a 1:2 chance of success of reaching someone who is planning on leaving to build their satisfaction.

References:

ProductPlan. (2021, August 12). *Churn*. Product Roadmap Software | ProductPlan. https://www.productplan.com/glossary/churn/

UMGC. (2023). *Churn at a Cellular Phone Company-mod3 Excel Spreadsheet*. Data-640 UMGC. https://learn.umgc.edu/d2l/le/content/920742/viewContent/31268436/View

Appendix

Name	ler	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
account_length		No	.	.	.	0	1	243	101.0648	39.82211	0.096606	-0.10784
area_code		No	.	.	3	0	.	.	.	.	.	.
churn		No	.	.	2	0	.	.	.	.	.	.
customer_service_calls		No	.	.	10	0	.	.	.	.	.	.
international_plan		No	.	.	2	0	.	.	.	.	.	.
number_vmail_messages		No	.	.	.	0	0	51	8.09901	13.68837	1.264824	-0.05113
phone_number		No	.	.	21	0	.	.	.	.	.	.
state		No	.	.	21	0	.	.	.	.	.	.
total_day_calls		No	.	.	.	0	0	165	100.4356	20.06908	-0.11179	0.243182
total_day_charge		No	.	.	.	0	0	59.64	30.56231	9.259435	-0.02908	-0.01981
total_day_minutes		No	.	.	.	0	0	350.8	179.7751	54.46739	-0.02908	-0.01994
total_eve_calls		No	.	.	.	0	0	170	100.1143	19.92263	-0.05556	0.206156
total_eve_charge		No	.	.	.	0	0	30.91	17.08354	4.310668	-0.02386	0.025487
total_eve_minutes		No	.	.	.	0	0	363.7	200.9803	50.71384	-0.02388	0.02563
total_intl_calls		No	.	.	.	0	0	20	4.479448	2.461214	1.321478	3.083589
total_intl_charge		No	.	.	.	0	0	5.4	2.764581	0.753773	-0.24529	0.60961
total_intl_minutes		No	.	.	.	0	0	20	10.23729	2.79184	-0.24514	0.609185
total_night_calls		No	.	.	.	0	33	175	100.1077	19.56861	0.0325	-0.07202
total_night_charge		No	.	.	.	0	1.04	17.77	9.039325	2.275873	0.008886	0.085663
total_night_minutes		No	.	.	.	0	23.2	395	200.872	50.57385	0.008921	0.085816
voice_mail_plan		No	.	.	2	0	.	.	.	.	.	.

Figure 1: Shows churn data set with all relevant variables.

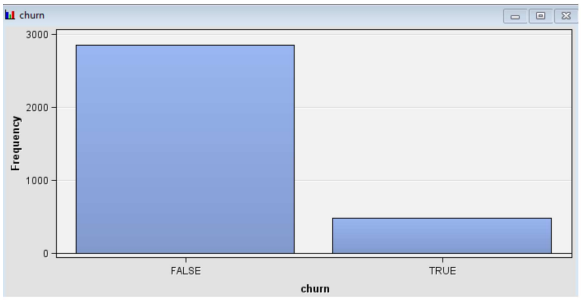

Figure 2: Bar chart of target variable Churn with 483 out of 3333 (14.5%) equal to True

Train	
Variables	...
Maximum Iterations	25
Use Missing as Level	No
Tolerance	1.0E-6
Penalty	1.0

Figure 3: Default training parameters used for each of the SVM models.

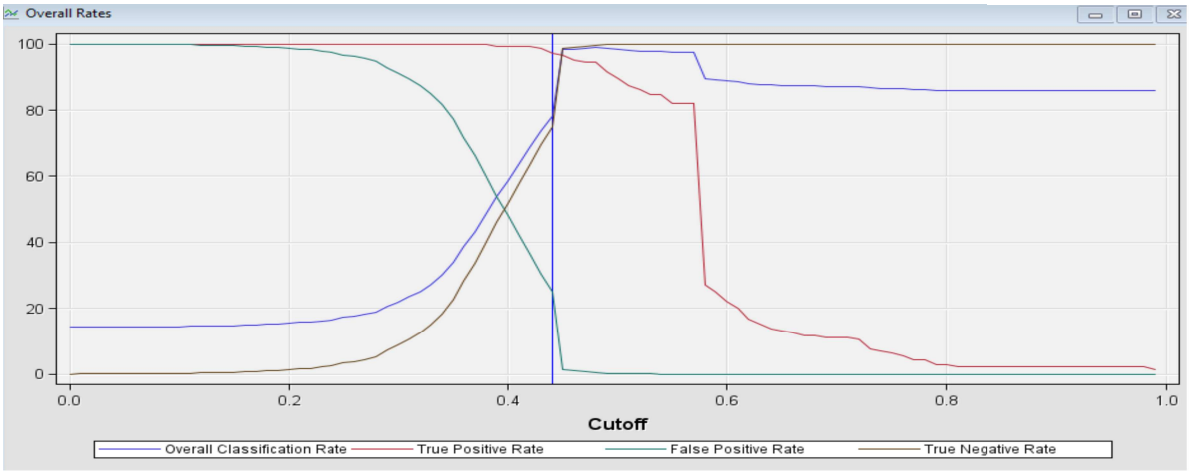

Figure 4: graph showing overall rates vs the cutoff value for the 3rd degree polynomial model. (0.44 optimal)

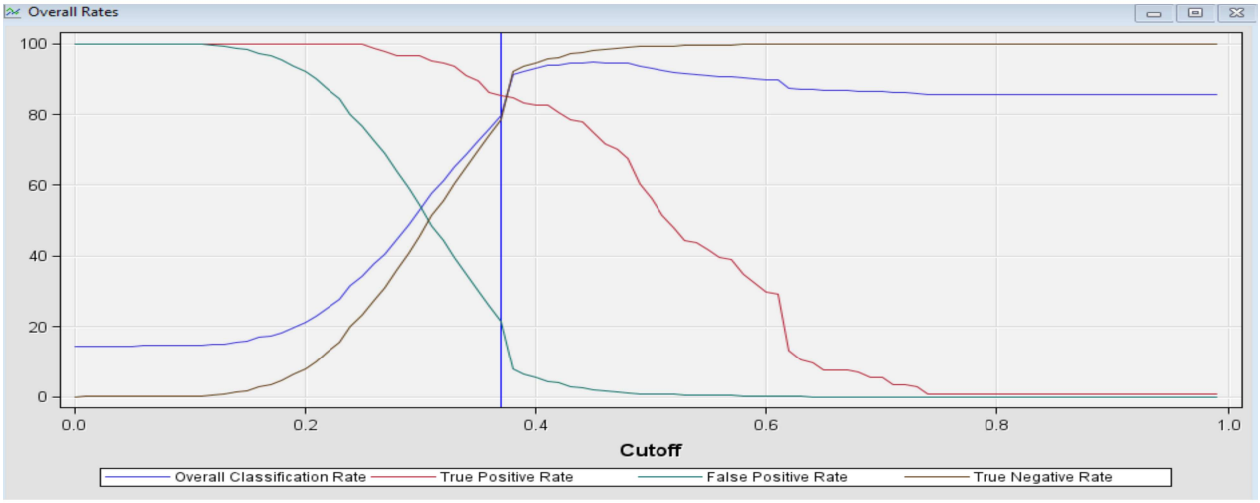

Figure 5: graph showing overall rates vs the cutoff value for the 2nd degree polynomial model. (0.37)

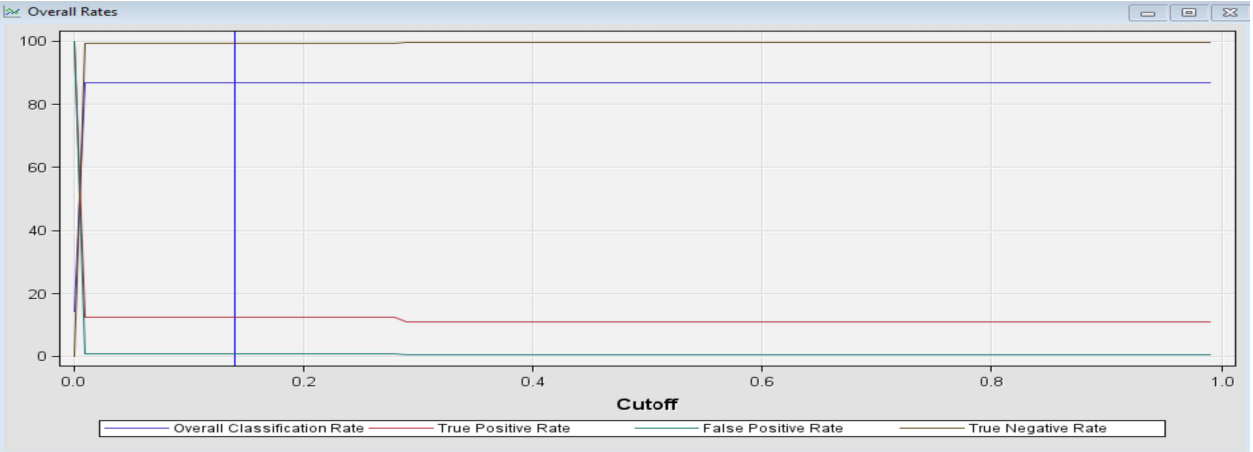

Figure 6: Graph showing overall rates vs the cutoff value for the Linear model. (0.02 optimal, but 0.14 was chosen)

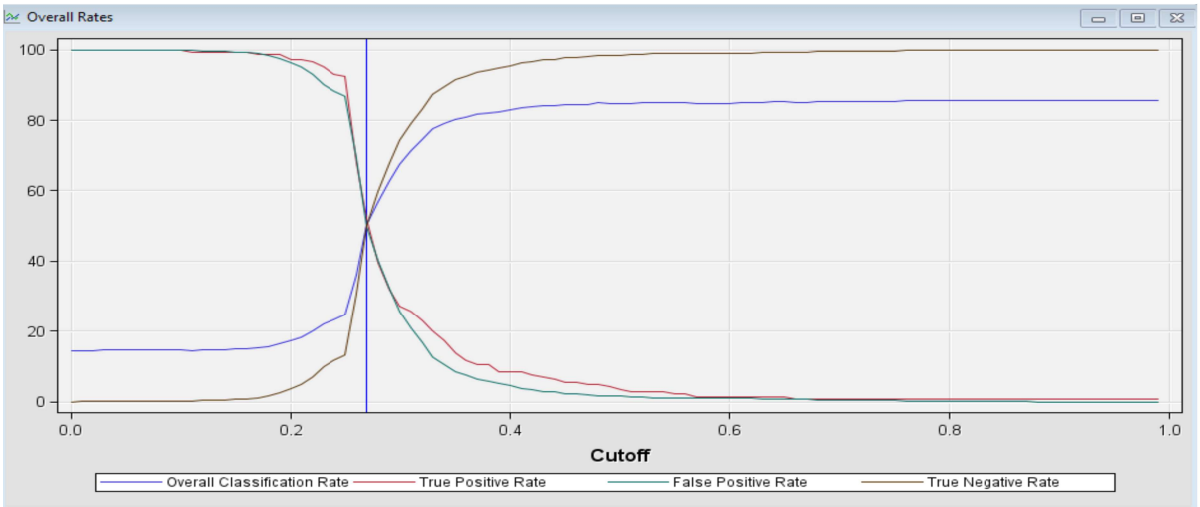

Figure 7: Graph showing overall rates vs the cutoff value for the Sigmoid +/- 1 model. (0.27 optimal)

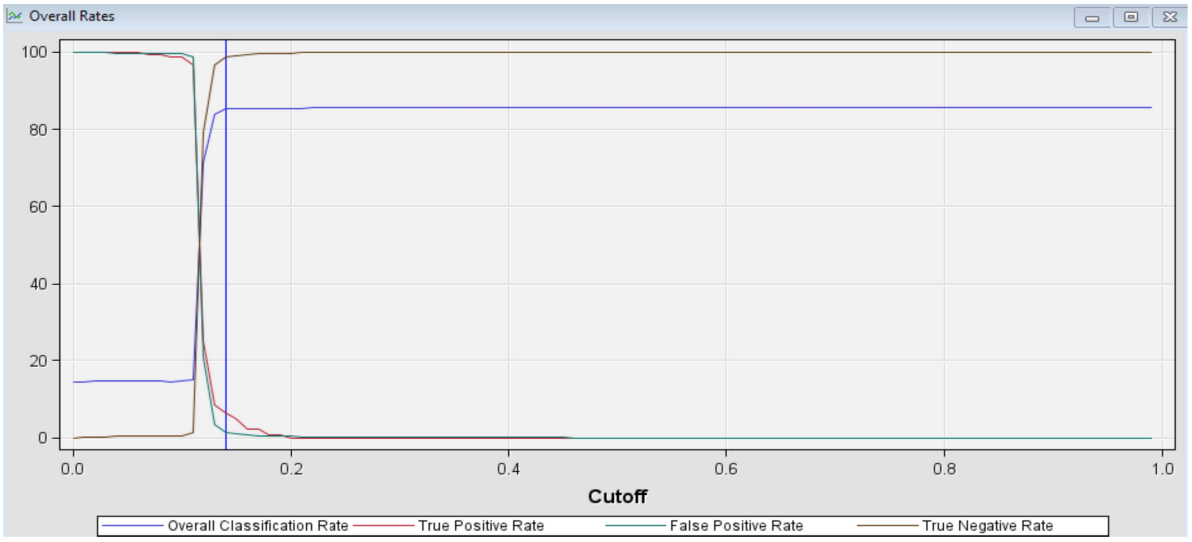

Figure 8: Graph showing overall rates vs the cutoff value for the Sigmoid +/- 2 model. (0.14 optimal)

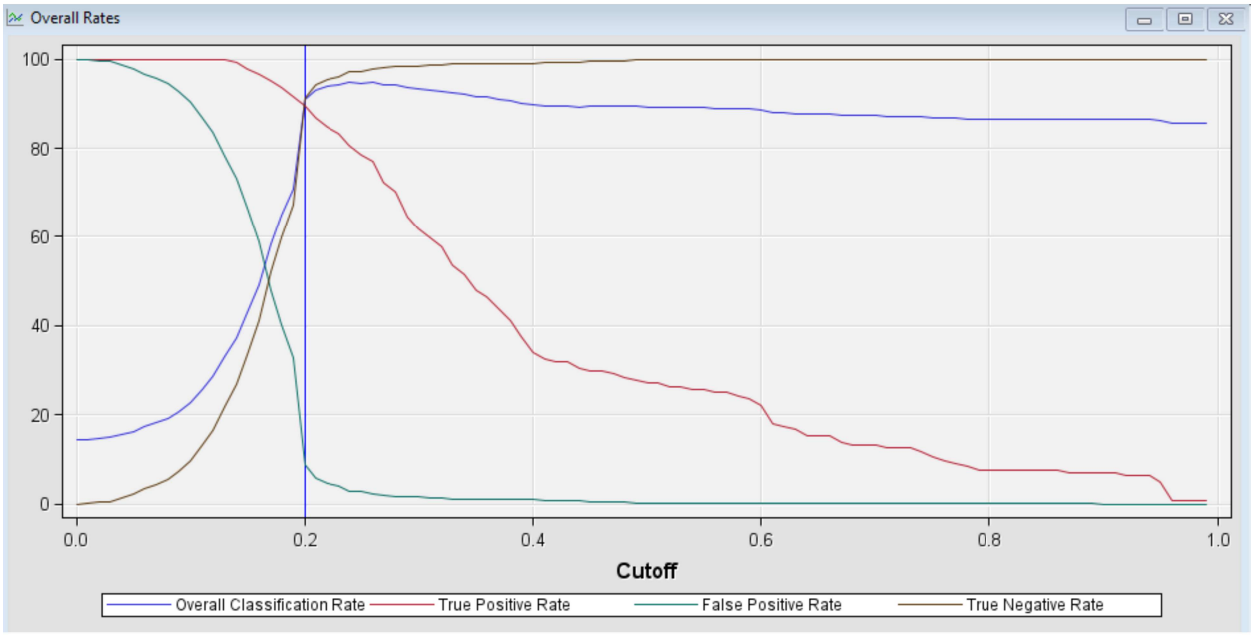

Figure 9: Graph showing overall rates vs the cutoff value for the RBF =1 model. (0.2 optimal)

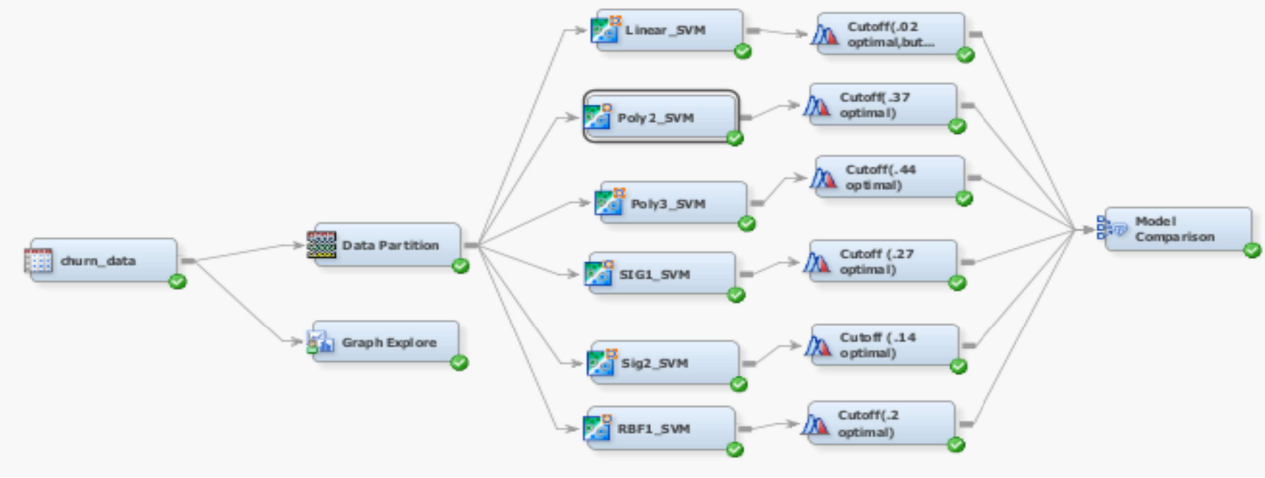

Figure 10: Full diagram of all 6 SVM models showing the original data, partition, models, and cutoff nodes.

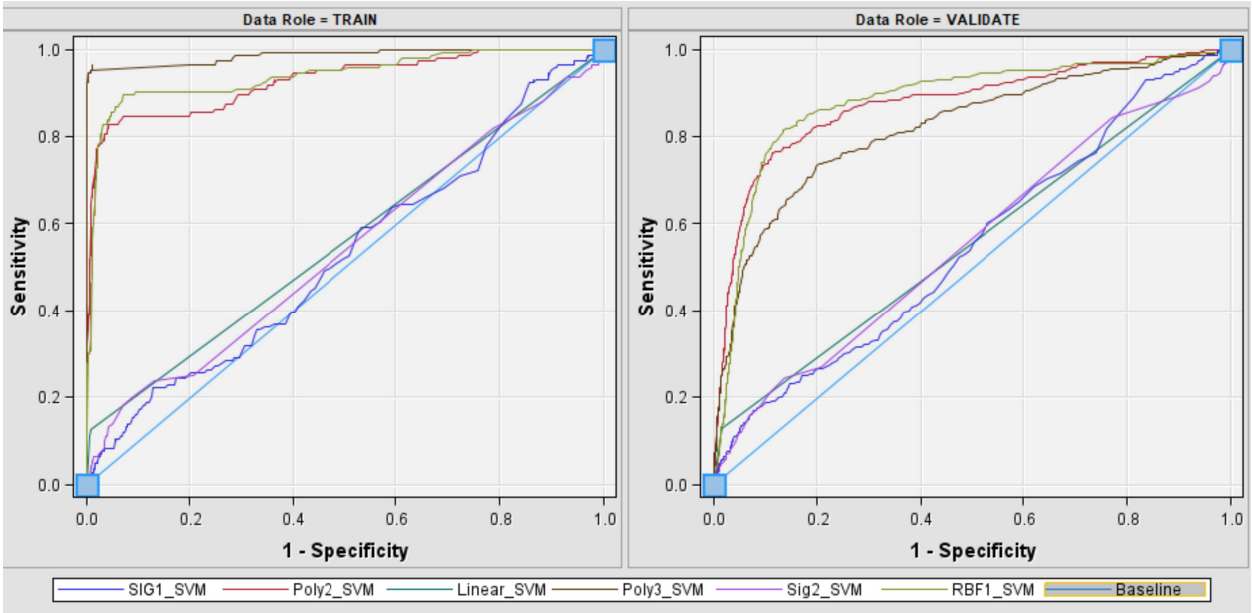

Figure 11: Shows the ROC curve for all 6 models, where the two polynomial models and RBF did the best.

Data Role=Valid

Statistics	HPSVM5	HPSVM3	HPSVM	HPSVM4	HPSVM2	HPSVM6
Valid: Kolmogorov-Smirnov Statistic	0.65	0.53	0.68	0.11	0.11	0.09
Valid: Average Squared Error	0.13	0.18	0.10	0.14	0.12	0.14
Valid: Roc Index	0.87	0.82	0.88	0.84	0.59	0.55
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.43	0.45	0.23	0.00	0.12	0.26
Valid: Cumulative Percent Captured Response	48.08	41.89	42.77	42.48	17.70	17.70
Valid: Percent Captured Response	23.01	16.81	22.71	21.53	7.96	6.78
Valid: Frequency of Classified Cases	2335.00	2335.00	2335.00	2335.00	2335.00	2335.00
Valid: Divisor for ASE	4670.00	4670.00	4670.00	4670.00	4670.00	4670.00
Valid: Gain	379.80	317.98	326.81	323.87	76.61	76.61
Valid: Gini Coefficient	0.75	0.64	0.76	0.68	0.19	0.10
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.64	0.52	0.67	0.57	0.20	0.09
Valid: Kolmogorov-Smirnov Probability Cutoff	0.41	0.45	0.23	0.00	0.12	0.36
Valid: Cumulative Lift	4.80	4.18	4.27	4.24	1.77	1.77
Valid: Lift	4.59	3.36	4.53	4.30	1.59	1.35
Valid: Maximum Absolute Error	0.87	0.81	0.97	1.00	0.99	0.90
Valid: Misclassification Rate	0.11	0.13	0.14	0.14	0.15	0.15
Valid: Sum of Frequencies	2335.00	2335.00	2335.00	2335.00	2335.00	2335.00
Valid: Root Average Squared Error	0.36	0.42	0.32	0.37	0.35	0.38
Valid: Cumulative Percent Response	69.66	60.68	61.97	61.54	25.64	25.64
Valid: Percent Response	66.67	48.72	65.81	62.39	23.08	19.66
Valid: Sum of Squared Errors	607.21	821.51	477.91	654.54	581.17	670.95
Valid: Number of Wrong Classifications	246.00	315.00	327.00	330.00	339.00	349.00

Figure 13: Shows statistical results of models, see figure 12 to understand which model label goes to which model type.