

Assignment 1: Support Vector Machine (SVM) Models Using SAS Enterprise Miner

Melissa Hunfalvay

Email: Melissa.Hunfalvay@gmail.com

Data 640 9040

Spring 2022

Professor Steve Knode

Date: February 1st, 2022

Introduction

The dataset chosen was “car lemon dataset.csv”. The data included various features of cars (such as color and make) and whether the cars were determined to be good or bad purchases. Using SAS Enterprise Miner software, the data lifecycle was conducted following the SEMMA (Sample, Explore, Modify, Model and Assess) methodology (SAS Institute, 1998). Support Vector Machines (SVM) models were used to examine the data.

The purpose of the analysis was to differentiate the bad vehicle purchases, from good vehicle purchases. More specifically:

- a) To look for key attributes that help to classify a vehicle as a “bad” purchase
- b) To buy more quality vehicles for the organization to sell
- c) To provide higher quality vehicles to customers leading to improved customer satisfaction

The car lemon dataset included 34 variables and 72,983 rows or observations. The type of variables included numeric ($n = 15$) and character ($n = 19$) variables. Of the numeric variables, 11 were interval, the other two were binary.

To accomplish the purpose of the analysis the variable “*Is Bad Buy*” was the identified as the dependent or target binary variable. “0” indicates that the purchase was not a bad buy, and “1” which indicates the purchase was a bad buy. The target variable was unevenly proportioned and skewed heavily in favor of “0” values or not bad purchases ($n = 64,007$; 87.70%). Bad purchases totaled 8,976 of the 72,983 observations (12.30%).

Data Cleaning and Preparation

The dataset was imported into SAS Enterprise Miner using the File Import node and converted to a .sas7bdat file using the Save Data node. Figure 1 shows the initial data import. Using the Sample Node, 27% of this data was used for the analysis ($n = 19,705$).

Several variables were rejected by SAS. To ensure these variables did not contribute meaningfully to understanding the data the Worth value in the Stat Explore node was examined (Figure 2). None of the rejected variables were displayed with a Worth score therefore these values remain rejected.

Using Edit Variables from the Data Source node the following nominal variables were rejected as they were identifying variables only: BYRNO, RefID, WheelTypeID. The *Is Bad Buy* variable was set as the target variable.

As recommended by Abbott (2014, p.85), a view of the raw data to determine if any obvious formatting errors were present was conducted and a check for incorrect values. Neither were found.

The next phase of understanding the data included a review of the descriptive statistics. The following checks of the data were undertaken a) missing data, b) large ranges in data (minimum and maximum values) which may indicate outliers c) skewness in data distributions, d) high cardinality of categorical variables, e) categorical variables with many single values, f) highly correlated variables (Abbott, 2014).

Missing data was found in five of the variables (Figure 3). These variables were imputed using default methods (Figure 4). Warranty Cost has the highest skewness value (2.07; Figure 5) and was transformed via a log 10 transformation (SAS Institute, 2015). Vehicle age (VehicleAge) and vehicle year (VehYear) were two measures of the same variable and were highly correlated. The time elapsed since the car was manufactured was deemed more appropriate than the year of manufacturing, therefore vehicle year was rejected. The final dataset after cleaning was complete is Figure 6. Figure 7 shows the diagram process flow to this point in the SEMMA methodology.

Predictive Model Development

A two-phased approach was used to develop the predictive models and determine the champion model. Phase one assessed parameters within each kernel algorithm and choose the “best within” kernel result. Phase 2: choose the best of all kernels, that is, the best overall model. Table 1 explains the predictive models developed and the parameters examined (see Figure 8 diagram process flow).

Table 1: Predictive Models Developed and Parameters Outlined

Kernel	Explanation	Algorithm	Models Developed: Parameters varied	
Linear	Separates the data using a hyperplane that is a line (Knode, 2022; James, Witten, Hastie, Tibshirani, 2017)	$K(u, v) = u^T v$	1	Cutoff threshold 0.36 (custom – best fit)
		Equation for a line. If left and right side of the equation are equal the point is on the line.	2	Cutoff Threshold 0.5 (default)
		K = Kernel function. u and v are vectors in the input space. T is the transpose of vector u .	3	Cutoff threshold 0.13 (imbalanced target percentage)
Polynomial	Is an algebraic function whereby the degree (order) can be set. If the order is in 3 dimensions, it is a “plane” if it is in more than 3 dimensions it is a hyperplane (Knode, 2022).		4	90/10 Data partition Cutoff threshold 0.13 (imbalanced target percentage) Polynomial Function 2
		$K(u, v) = (u^T v + 1)^p$	5	70/30 Data Partition Cutoff threshold 0.13 (imbalanced target percentage) Polynomial Function 2
		p is the order of the polynomial. K = Kernel function.	6	70/30 Data Partition Cutoff threshold 0.32 (custom – best fit) Polynomial Function 2
		1 is added to avoid zero value entries. T is the transpose of vector u . u and v are vectors in the input space.	7	70/30 Data Partition Cutoff threshold 0.13 (imbalanced target percentage) Polynomial Function 2
			8	70/30 Data Partition Cutoff threshold 0.44 (custom) Polynomial Function 3
Radial Based Function (RBF)	A spherical (circular) function where any of the line segments from a central point to the perimeter (Abbey, He & Wang, n.d.).		9	70/30 Data Partition Cutoff threshold 0.24 (custom – best fit) RBF Degree: 3
		$K(u, v) = \exp[-p(u - v)^2]$	10	70/30 Data Partition Cutoff threshold 0.47 (custom – best fit) RBF Degree: 1
		\exp is the exponential function K = Kernel function.	11	90/10 Data partition Cutoff threshold 0.13 (imbalanced target percentage) RBF Degree: 1
		T is the transpose of vector u . u and v are vectors in the input space.	12	70/30 Data Partition Cutoff threshold 0.13 (imbalanced target percentage) RBF Degree: 1
			13	70/30 Data Partition Cutoff threshold 0.13 (imbalanced target percentage) RBF Degree: 3
Sigmoid Function	Sigmoid is an “S” shaped curve. Sigmoid 1 specifies the amount of scaling inside the sigmoid function. Whereas Sigmoid 2 specifies the amount of <i>bias</i> inside the sigmoid function	$K(u, v) = \tanh(p(u^T v) + q)$	14	Sigmoid 1: 3 degrees of scaling
		K = Kernel function. u and v are vectors in the input space.	15	Sigmoid 1: 1 degree of scaling
		Where p is the kernel scale and q is the kernel location parameter, \tanh is the hyperbolic tangent function	16	Sigmoid 2: -1 degree of bias
		T is the transpose of vector u .	17	Sigmoid 2: -5 degrees of bias

To develop the predictive models, data needed to be partitioned to provide mutually exclusive data sets for comparison. Using the Data Partition Node, data was divided into training and validation sets. The purpose of the training set is for model fitting, whereby model parameters are examined to determine the best fit before a comparison of the models (SAS Enterprise Miner, n.d.). The validation set is used to check for an unbiased estimate of the generalizability of the model to determine if overfitting was occurring (Abbott, 2014).

A total of 17 models were developed and compared (see Table 2). All available kernels were examined. Parameters within kernels varied, for example, sigmoid functions changed in degree of bias and degrees of scaling.

Imbalanced Target

The binary target variable “*Is Bad Buy*” was imbalanced. An adjustment was needed to ensure enough of the bad buys were included in the model. This is especially important as the purpose of the analysis is to identify these rare events and to do so the models need enough of these events to provide meaningful findings.

The cutoff criterion using the Cutoff Node was implemented. The initial threshold used was changed from the default of 0.5 to 0.13 to reflect the proportion of the bad buys within the overall dataset as recommended by Abbott (2014). Once results were examined, the table in the Cutoff Node guided further iterations of the cutoff thresholds to find the best overall model (see Table 2). The cutoff method provided a more accurate model for identification of the true positives, that is real data and model data agreement on bad buys (or 1 values). The disadvantage of this method is the likelihood of also identifying many more false positives (i.e. real data that is a good buy being identified as a bad buy).

As a counterbalance, the cost function was also calculated to determine overall cost of each model and specifically, the cost of false positive results (see Table 2). Costs for true values (negative and positive) were set to \$0, cost for false negatives was \$6.90 per instance and for false positives was \$1.00.

Accuracy Measures and Results

Table 2 provides an assessment of each model. The purpose of this assessment was to identify rare (positive or “1”) events, which are bad buys. Therefore, models were assessed in the following order of importance to achieve this goal. First, identification of true positives, that is, bad buys. Second, sensitivity, the models ability to identify positives that were actually positives in the real data. Third, precision (or the Positive Predictive Value), which is the models ability to identify true positive events among all positive events. Fourth, F1 score, which is the weighted average of precision and sensitivity. Fifth, accuracy, which is the models ability to accurately classify all the events in the real data.

All these measures were examined in both training and validation data to determine the ability of the model to generalize (see “Delta” column in Table 2, which is a calculation of the absolute difference in accuracy between the training and validation scores).

When comparing models several other factors were considered (see Table 3). First, Receiver Operating Curves (ROC), were used in the Model Comparison Node to visually compare the tradeoff between sensitivity and specificity. An ROC index of 0.9 or greater is considered an excellent model (Adjorlolo, 2018). Second, cumulative lift, which indicates the models ability to predict beyond chance (Figure 9). A lift of greater than 2 is considered significantly improved compared to no model. Third, the cost of the model. Fourth, the misclassification rate.

Table 2: Model Comparisons Accuracy Measures

Model Characteristics				Training Dataset										Validation Dataset					Notes	Conclusion	
Model #	Data Partition	Cutoff	Kernel	Value	# TP	#FN	Sensitivity (%)	Accuracy (%)	Precision (%)	F1	#TN	#FP	Specificity (%)	Total Cost	#TP	#FN	Sensitivity (%)	Accuracy (%)	Delta		
1	70/30	0.36	Linear	N/A	1693	3	99.82	14.19	55.69	0.222	265	11831	0.02	\$11,852	709	18	0.98	14.44	0.25	High sensitivity. High true positives. Trade off: High number of false positives whereby model predicts a bad buy that is not accurate.	Accept: Phase 2
9	70/30	0.25	RBF	3	1672	24	98.58	34.99	83.65	0.272	3155	8941	0.26	\$9,107	614	113	0.84	33.65	1.34	Good identification of True positives. False negatives low. High sensitivity. High precision. Low-moderate accuracy. Second lowest cost. Better specificity compared with other models in phase 2 (except #10). Overall champion?	Accept: Phase 2
16	70/30	0.13	Sigmoid 2	-1	1696	0	100	12.35	56.15	0.219	8	12088	0.00	\$12,088	727	0	1.00	12.32	0.03	Good sensitivity. Accuracy very low due to low cut off threshold. Tradeoff high false positives. F1 value low. High cost.	Accept: Phase 2
6	70/30	0.32	Polynomial	2	1695	1	99.94	13.57	55.94	0.221	177	11919	0.01	\$11,926	714	13	0.98	13.74	0.17	Good sensitivity. Moderate precision. Low accuracy. High cost.	Accept: Phase 2
10	70/30	0.47	RBF	1	730	966	43.04	92.9	94.77	0.599	12084	12	1.00	\$6,677	147	580	0.20	89.51	3.39	Not enough identification of TP leading to lowest sensitivity in phase 2. Good accuracy, Good precision. Best F1 score. Lowest overall cost. Overall champion?	Accept: Phase 2
14	70/30	0.13	Sigmoid 1	3	1696	0	100	12.35	56.15	0.219	8	12088	0.00	\$12,088	727	0	1.00	12.32	0.03	Good sensitivity. Accuracy very low due to low cut off threshold. Tradeoff high false positives. F1 value low. High cost.	Accept: Phase 2
2	70/30	0.5	Linear	N/A	505	1191	29.77	90.82	89.1	0.444	12022	74	0.99	\$8,292	168	559	0.23	88.99	1.83	Low sensitivity. Not able to identify bad buys.	Reject
3	70/30	0.13	Linear	N/A	1695	1	99.94	12.52	54.68	0.219	33	12063	0.00	\$12,070	727	0	1.00	12.53	0.01	Change in cutoff threshold significantly improved sensitivity. Tradeoff: more true positives but also more false positives. Increase in cost.	Reject
4	90/10	0.13	Polynomial	2	2180	0	100	12.34	56.14	0.219	9	15544	0.00	\$15,544	243	0	1.00	12.32	0.02	High number of false positives because low cut-off threshold and 90/10 data split added instances to validation dataset. Cost highest of all models.	Reject
5	70/30	0.13	Polynomial	2	1696	0	100	12.31	56.14	0.219	2	12094	0.00	\$12,094	727	0	1.00	12.29	0.02	High number of false positives because low cut-off threshold	Reject
17	70/30	0.13	Sigmoid 2	-5	1696	0	100	12.3	56.14	0.219	1	12095	0.00	\$12,095	727	0	1.00	12.36	0.06	High number of false positives because low cut-off threshold	Reject
7	70/30	0.13	Polynomial	2	1696	0	100	12.34	56.15	0.219	6	12095	0.00	\$12,095	727	0	1.00	12.29	0.05	High number of false positives because low cut-off threshold	Reject
8	70/30	0.44	Polynomial	2	569	1127	33.54	90.82	85.87	0.473	11975	139	0.99	\$7,915	207	520	0.28	88.28	2.54	Low sensitivity. Not able to identify bad buys.	Reject
11	90/10	0.13	RBF	1	2011	169	92.24	97.58	93.74	0.904	15294	259	0.98	\$1,425	138	105	0.57	64.19	33.39	Overfit (unfortunately). When examining cutoff thresholds throughout the range from 0.1 to 1.0 significant overfitting occurred throughout. Validation model great. Training model did not generalize.	Reject
12	70/30	0.13	RBF	1	1562	134	92.09	97.73	94.32	0.909	11918	178	0.99	\$1,103	432	295	0.59	65	32.73	Same as above	Reject
13	70/30	0.13	RBF	3	1695	1	99.94	12.53	54.68	0.219	33	12063	0.00	\$12,070	727	0	1.00	12.49	0.04	High number of false positives because low cut-off threshold	Reject

Table 3: Model Comparison Results and Insights

Model #	Training			Validation		
	Name in Diagram	ROC Index	Misclassification Rate	Cumulative Lift	ROC Index	Misclassification Rate
1	Linear 0.36	0.75	0.11	3.77	0.75	0.11
9	RBF	0.92	0.09	6.67	0.69	0.11
16	Sigmoid 2	0.55	0.2	1.05	0.58	0.21
6	Active Poly 2	0.78	0.09	5.02	0.67	0.11
10	RBF 47	0.96	0.11	7.14	0.69	0.1
14	Sigmoid 1	0.54	0.12	1.36	0.55	0.12

Conclusion and Takeaways

The goal of this analysis was to determine the bad buys (true positives) within the car dataset. Therefore, determination of the “champion” model was between model 9 (RBF) and model 10 (RBF47). The champion model needed to identify the greatest number of bad buys (model 9, $n = 1672$; model 10, $n = 730$); high sensitivity (model 9: 98.58%; model 10: 43.04%); high precision, (model 9: 83.65%; model 10: 94.77%). Model 9 was able to achieve the goal of the analysis by identifying bad buys while balancing the tradeoff with false positives. Therefore, the overall “champion” model was model 9 (RBF).

Both models (9 & 10) produced validation results with very high cumulative lift rates (model 9: 3.67; model 10: 3.55; Figure 9, Table 3), and very low misclassification rates (Table 3, Figures 12 & 13). The overall cost for model 9 was \$9,107 and for model 10 was \$6,677. Validation ROC Indexes boarded on fair results (Adjorlolo, 2018; Figure 10, 11).

A caveat of the analysis is the use of a relatively small sample size. This may have been mitigated with a larger dataset especially for model 10 and may help to explain the reduced ROC indexes from training to validation test sets.

Limitations include the use of zip code (VNZIP1) as a number rather than a recommended text variable (Abbott, 2014, p.110). The actual zip code number is rarely the “causal reason for the behavior being measured, but rather a characteristic about the people who live in the [region] zip code.” A disadvantage of the dataset was the imbalanced set of target variables, ideally, more “bad buys” would help to classify this outcome. In summation, the car sales company has a model that will help identify the bad car sales, helping to improve inventory which will lead to happier customers and may lead to renewal purchases and referrals (it’s always more difficult to get a customer for the first time, than to retain them for purchases over time).

Appendix

Name	Role	Level	Type	Report	Order	Drop	Lower Limit	Upper Limit	Format	Informat	Length
AUCGUART	Input	Nominal	Character	No		No	.	.	\$5.	\$5.	5
Auction	Input	Nominal	Character	No		No	.	.	\$5.	\$5.	5
BYRNO	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
Color	Input	Nominal	Character	No		No	.	.	\$6.	\$6.	6
IsBadBuy	Target	Binary	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
IsOnlineSale	Input	Binary	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
MMRAcquisitionAuctionAveragePric	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
MMRAcquisitionAuctionCleanPrice	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
MMRAcquisitionRetailAveragePrice	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
MMRAcquisitionRetailCleanPrice	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
MMRCurrentAuctionAveragePrice	Rejected	Nominal	Character	No		No	.	.	\$5.	\$5.	5
MMRCurrentAuctionCleanPrice	Rejected	Nominal	Character	No		No	.	.	\$5.	\$5.	5
MMRCurrentRetailAveragePrice	Rejected	Nominal	Character	No		No	.	.	\$5.	\$5.	5
MMRCurrentRetailCleanPrice	Rejected	Nominal	Character	No		No	.	.	\$5.	\$5.	5
Make	Rejected	Nominal	Character	No		No	.	.	\$10.	\$10.	10
Model	Rejected	Nominal	Character	No		No	.	.	\$20.	\$20.	20
Nationality	Input	Nominal	Character	No		No	.	.	\$14.	\$14.	14
PRIMEUNIT	Input	Nominal	Character	No		No	.	.	\$4.	\$4.	4
PurchDate	Time ID	Interval	Numeric	No		No	.	.	MMDDYY10.0	MMDDYY10.0	8
RefId	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
Size	Input	Nominal	Character	No		No	.	.	\$11.	\$11.	11
SubModel	Rejected	Nominal	Character	No		No	.	.	\$30.	\$30.	30
TopThreeAmericanName	Input	Nominal	Character	No		No	.	.	\$8.	\$8.	8
Transmission	Input	Nominal	Character	No		No	.	.	\$6.	\$6.	6
Trim	Rejected	Nominal	Character	No		No	.	.	\$3.	\$3.	3
VNST	Rejected	Nominal	Character	No		No	.	.	\$2.	\$2.	2
VNZIP1	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
VehBCost	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
VehOdo	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
VehYear	Rejected	Nominal	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
VehicleAge	Input	Nominal	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
WarrantyCost	Input	Interval	Numeric	No		No	.	.	BEST12.0	BEST32.0	8
WheelType	Input	Nominal	Character	No		No	.	.	\$7.	\$7.	7
WheelTypeID	Input	Nominal	Character	No		No	.	.	\$4.	\$4.	4

Figure 1: Variables after initial import of data to SAS Enterprise Miner

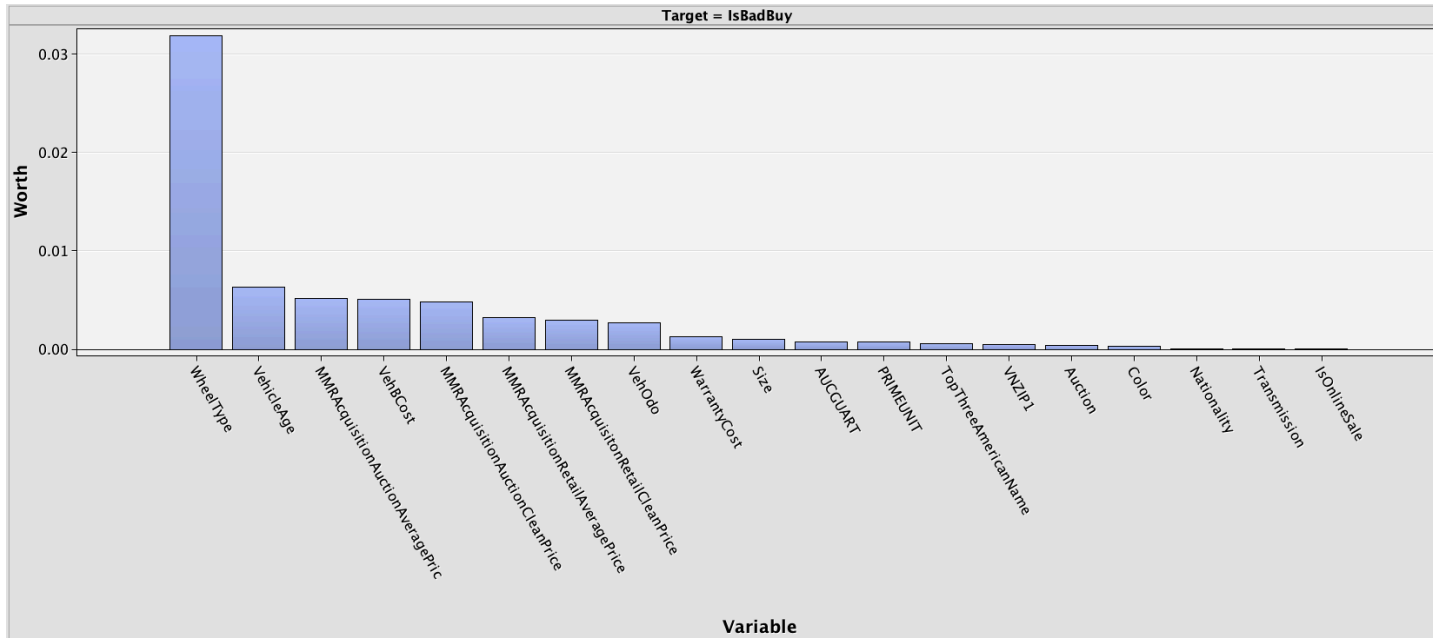


Figure 2: Worth value of the initial dataset imported via SAS.

Name	Percent Missing	Role
VNZIP1	0	Input
BYRNO	0	Rejected
MMRCurrentRetailAveragePrice	0	Rejected
RefId	0	Rejected
Color	0	Input
Model	0	Rejected
TopThreeAmericanName	0	Input
VehYear	0	Rejected
WheelTypeID	0	Rejected
MMRCurrentAuctionAveragePrice	0	Rejected
MMRCurrentRetailCleanPrice	0	Rejected
SubModel	0	Rejected
Make	0	Rejected
MMRCurrentAuctionCleanPrice	0	Rejected
PRIMEUNIT	0	Input
Nationality	0	Input
WarrantyCost	0	Input
IsOnlineSale	0	Input
VehicleAge	0	Input
AUCGUART	0	Input
Size	0	Input
Auction	0	Input
IsBadBuy	0	Target
VNST	0	Rejected
WheelType	0	Input
VehBCost	0	Input
VehOdo	0	Input
Transmission	0.00137	Input
MMRAcquisitionRetailCleanPrice	0.024663	Input
MMRAcquisitionRetailAveragePrice	0.024663	Input
MMRAcquisitionAuctionAveragePrice	0.024663	Input
MMRAcquisitionAuctionCleanPrice	0.024663	Input
Trim	4.367863	Rejected

Figure 3: Missing values in the data set

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
MMRAcquisitionAuctionAveragePrice	MEAN	IMP MMRACquisitionA...	6128.9092167	INPUT	INTERVAL		18
MMRAcquisitionAuctionCleanPrice	MEAN	IMP MMRACquisitionA...	7373.636031	INPUT	INTERVAL		18
MMRAcquisitionRetailAveragePrice	MEAN	IMP MMRACquisitionR...	8497.0343315	INPUT	INTERVAL		18
MMRAcquisitionRetailCleanPrice	MEAN	IMP MMRACquisitionRe...	9850.9282396	INPUT	INTERVAL		18
REP Color	COUNT	IMP REP Color	SILVER	INPUT	NOMINAL	Replacement: Color	8
REP Nationality	COUNT	IMP REP Nationality	AMERICAN	INPUT	NOMINAL	Replacement: Nationality	5
REP Size	COUNT	IMP REP Size	MEDIUM	INPUT	NOMINAL	Replacement: Size	5
REP TopThreeAmericanName	COUNT	IMP REP TopThreeAm...	GM	INPUT	NOMINAL	Replacement: TopThre...	5
REP Transmission	COUNT	IMP REP Transmission	AUTO	INPUT	NOMINAL	Replacement: Transmi...	9
REP WheelType	COUNT	IMP REP WheelType	Alloy	INPUT	NOMINAL	Replacement: WheelType	3174

Figure 4: Imputed values for the missing variables

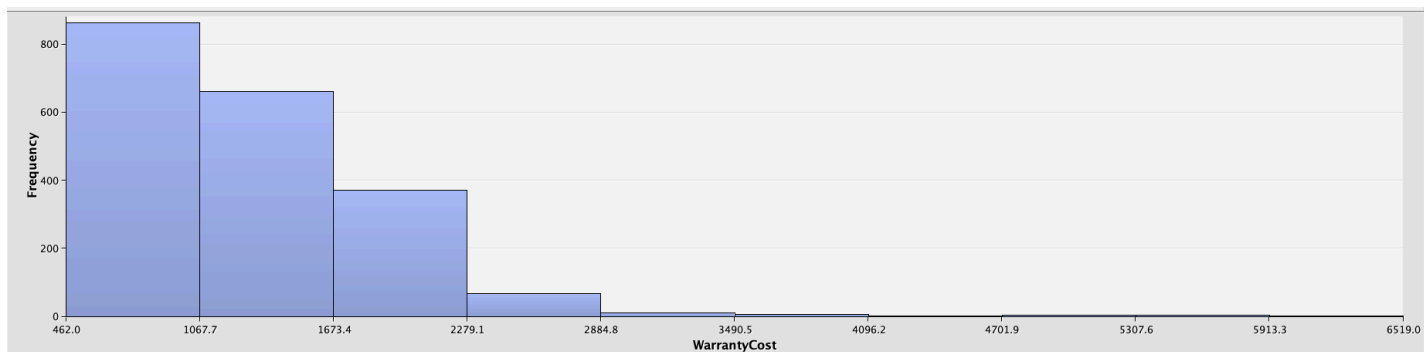


Figure 5: Skewness in the Warranty Cost variable shown using the replacement node


Name	Role	Level 	Minimum	Maximum	Mean	Standard Deviation
IsBadBuy	Target	Binary
IsOnlineSale	Input	Binary
MMRAcquisitionRetailCleanPrice	Input	Interval	0	41482	9850.928	3385.79
RefId	Rejected	Interval	1	73014	36511.43	21077.24
MMRAcquisitionAuctionCleanPrice	Input	Interval	0	36859	7373.636	2722.492
MMRAcquisitionRetailAveragePrice	Input	Interval	0	39080	8497.034	3156.285
VehOdo	Input	Interval	4825	115717	71500	14578.91
VehBCost	Input	Interval	1	45469	6730.934	1767.846
PurchDate	Time ID	Interval
WarrantyCost	Input	Interval	462	7498	1276.581	598.8468
BYRNO	Rejected	Interval	835	99761	26345.84	25717.35
VNZIP1	Input	Interval	2764	99224	58043.06	26151.64
MMRAcquisitionAuctionAveragePric	Input	Interval	0	35722	6128.909	2461.993
Transmission	Input	Nominal
WheelType	Input	Nominal
TopThreeAmericanName	Input	Nominal
SubModel	Rejected	Nominal
WheelTypeID	Rejected	Nominal
VehYear	Rejected	Nominal
VNST	Rejected	Nominal
Trim	Rejected	Nominal
VehicleAge	Input	Nominal
MMRCurrentAuctionAveragePrice	Rejected	Nominal
Make	Rejected	Nominal
MMRCurrentAuctionCleanPrice	Rejected	Nominal
Auction	Input	Nominal
AUCGUART	Input	Nominal
Color	Input	Nominal
PRIMEUNIT	Input	Nominal
Nationality	Input	Nominal
Size	Input	Nominal
MMRCurrentRetailCleanPrice	Rejected	Nominal
MMRCurrentRetailAveragePrice	Rejected	Nominal
Model	Rejected	Nominal

Figure 6: Final Data set after cleaning

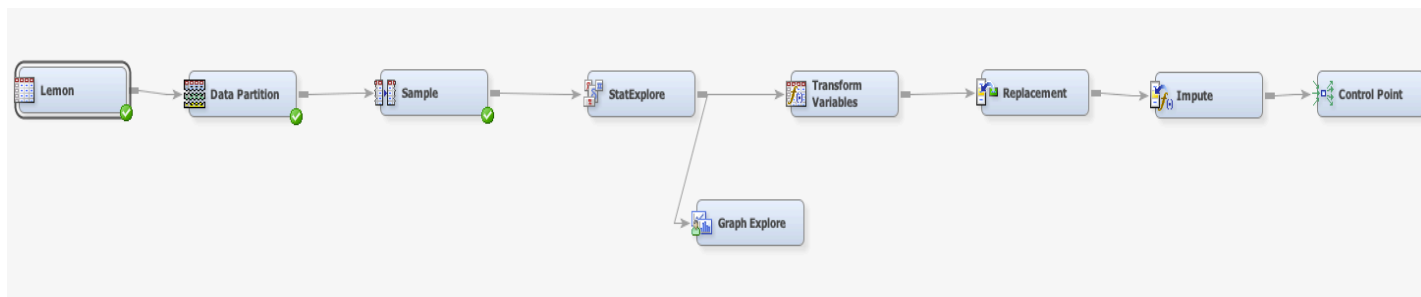


Figure 7: Diagram process flow SEMMA methodology Sample, Explore, Modify phases

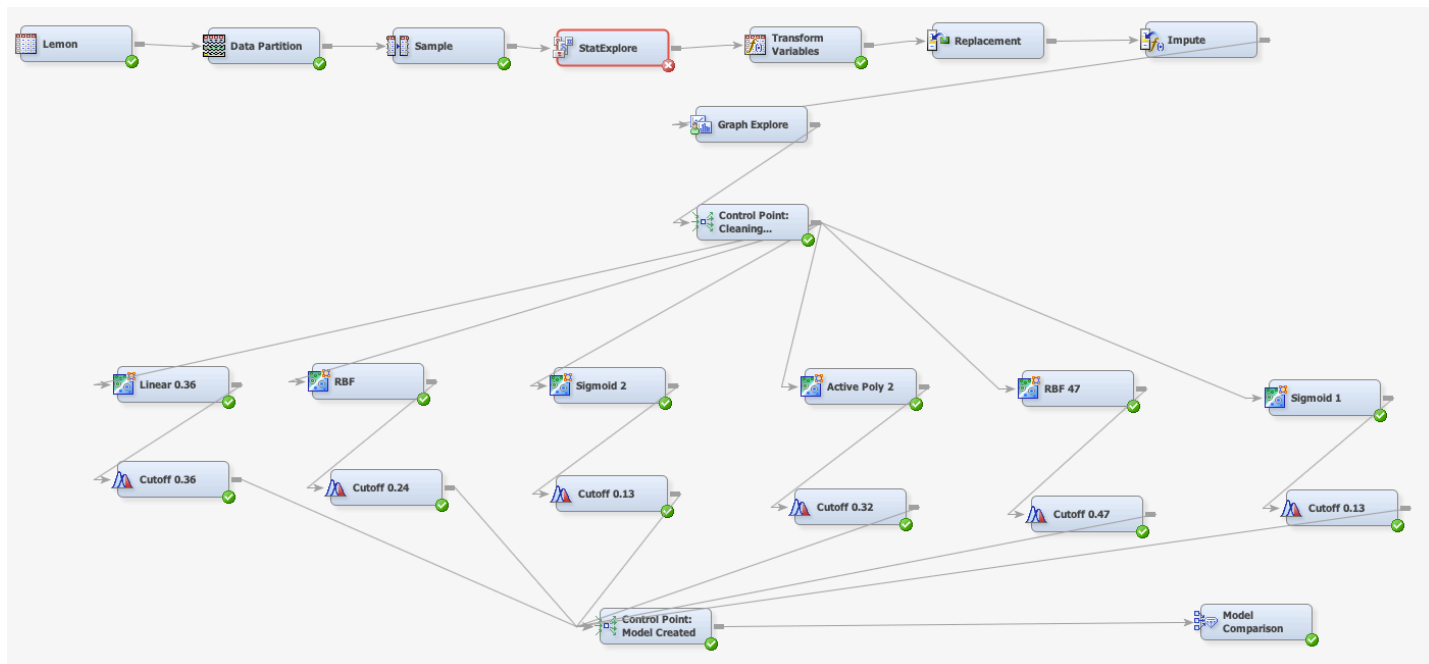


Figure 8: Diagram process flow for entire SEMMA methodology, including model comparison

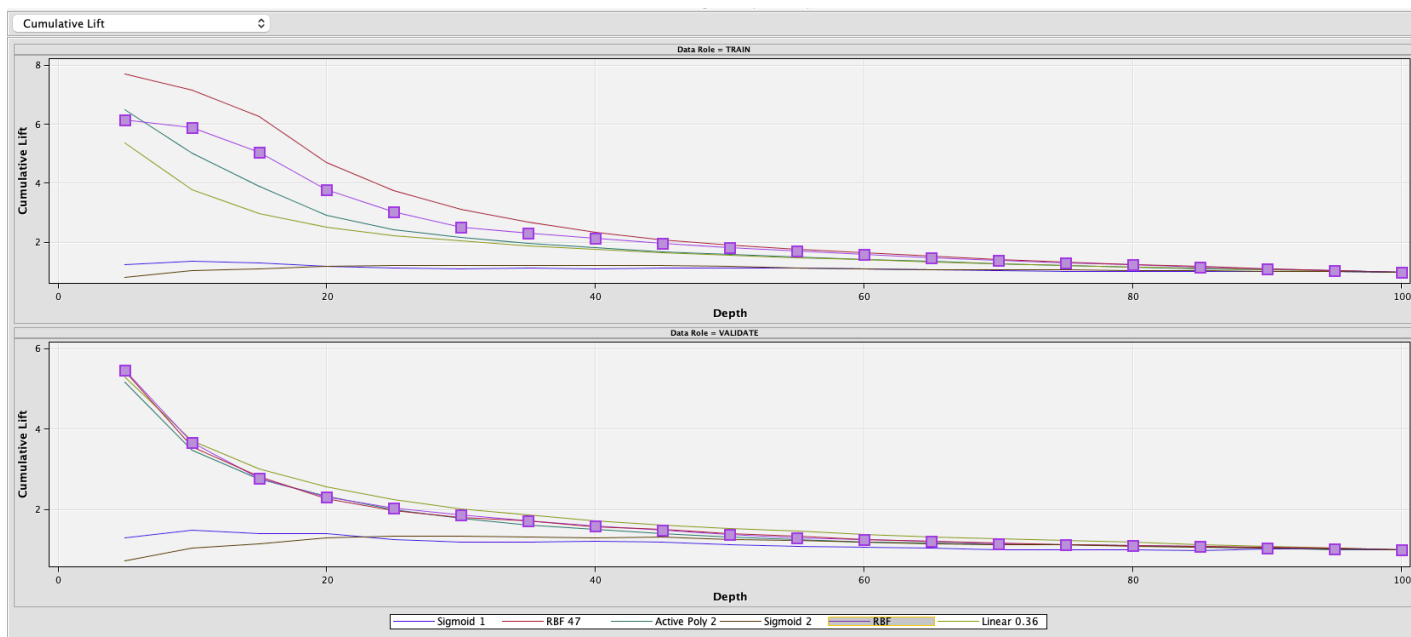


Figure 9: Cumulative lift: RBF model 9 in pink.

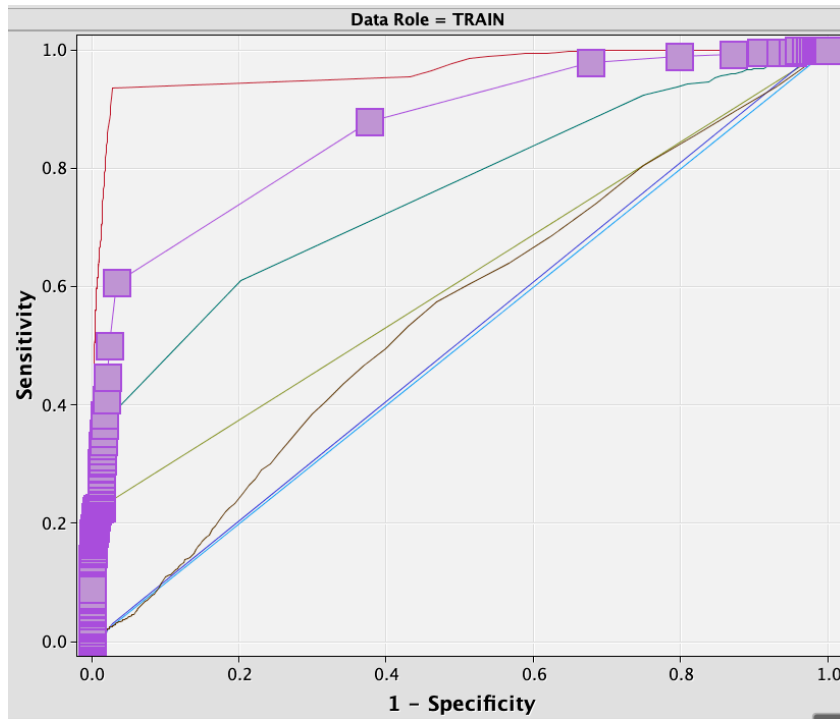


Figure 10: ROC curve for training dataset. Pink is the RBF model 9

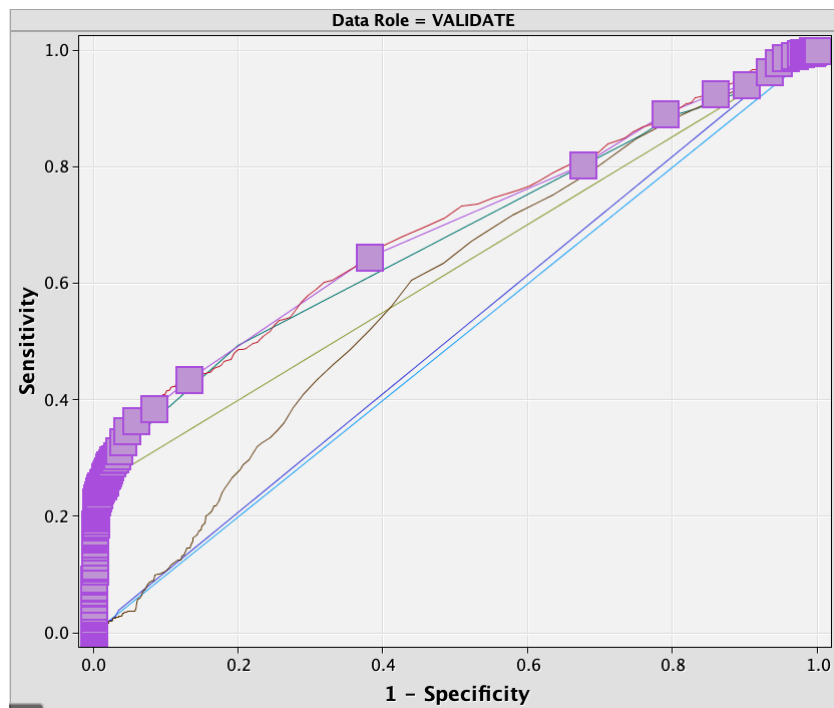


Figure 11: ROC curve for the validation training data set. In pink is model 9, RBF

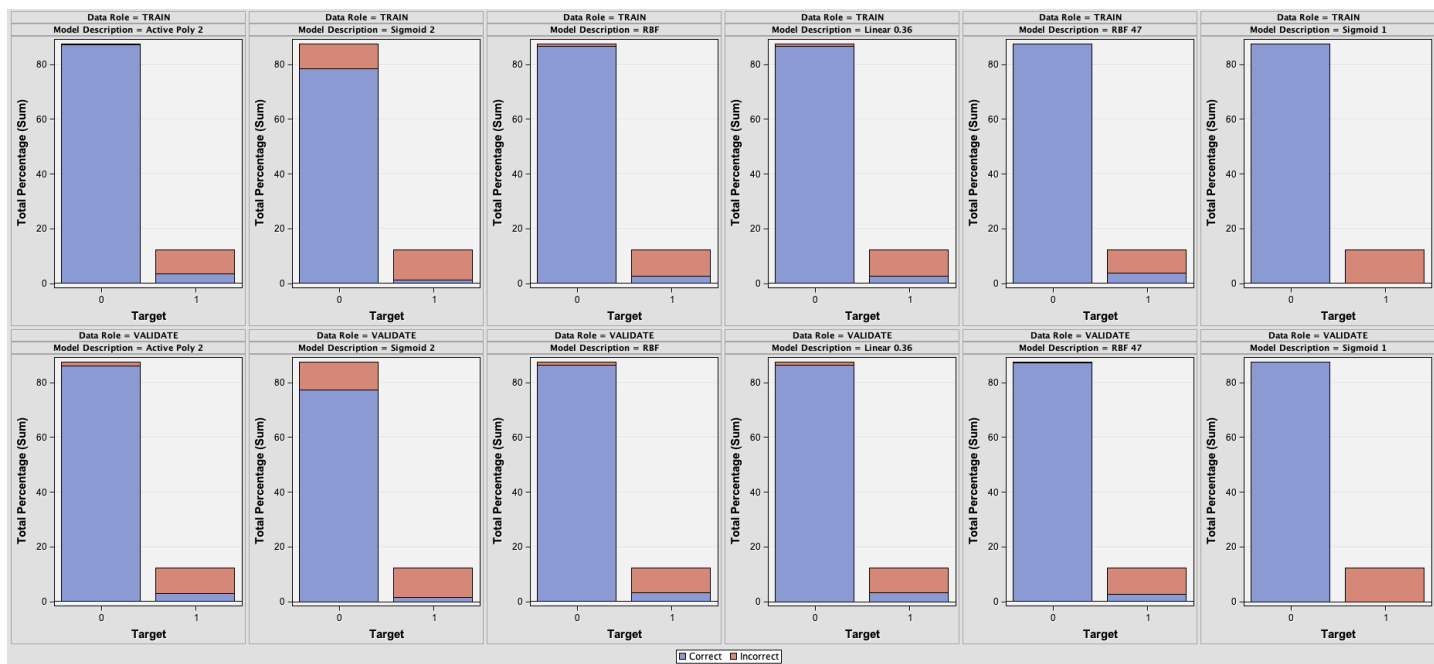


Figure 12: Misclassification Charts

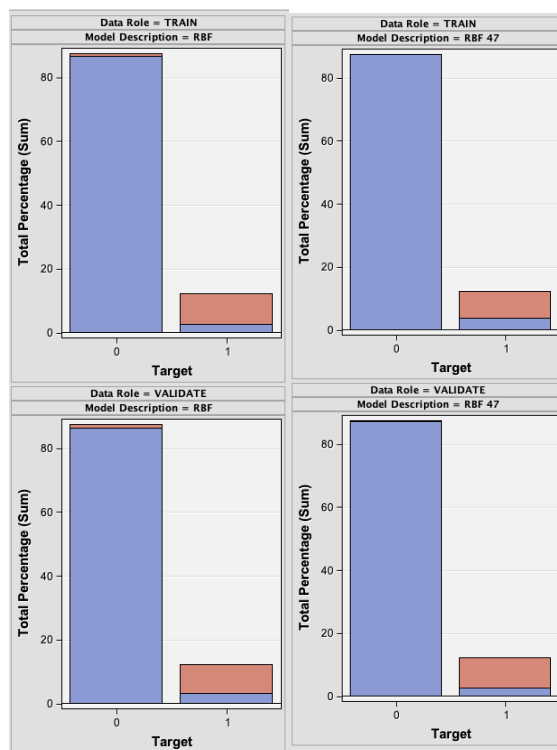


Figure 13: Misclassification events for the two champion model candidates

References

- Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Indianapolis, IN: Wiley Publishing.
- Adjorlolo, S. (2018). Diagnostic accuracy, sensitivity, and specificity of executive function tests in moderate traumatic brain injury in Ghana. *Assessment*, 25, 498–512. DOI: 10.1177/1073191116646445
- Buhmann, M. (2010) Radial basis function. *Scholarpedia*, 5(5): 9837.
- Gertz, E. M., and Griffin, J. D. (2005). *Support Vector Machine Classifiers for Large Data Sets*. Technical Report ANL/MCS-TM-289, Mathematics and Computer Science Division, Argonne National Laboratory.
- Kane, D. (2015). *Data science part IX: Support Vector Machine* [Web]. Retrieved October 24, 2017, from <https://www.youtube.com/watch?v=fMWjhQ2UcNs>
- Ray, S. (2017). *Understanding Support Vector Machine algorithm from examples (along with code)*. Retrieved October 23, 2017, from <https://www.analyticsvidhya.com/>
- SAS Institute Inc. (1998). *SAS Institute White Paper: Data Mining and the Case for Sampling*. Cary, NC: SAS Institute Inc. Retrieved from: https://scweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf
- SAS Institute Inc. (2015). *Getting Started with SAS Enterprise Miner: Imputing and Transforming Data*. Cary, NC: SAS Institute Inc. Retrieved from: <https://www.youtube.com/watch?v=TnWRJQb5z4c>
- SAS Enterprise Miner (n.d.). *Data Partition Node*. Cary, NC: SAS Institute Inc. Retrieved from: <https://documentation.sas.com/doc/en/emref/15.1/n0u3s4tv5v88cfn1dx0zkfd9pjm5.htm>