Title Page

Assignment 6: Automated Machine Learning

Melissa Hunfalvay

Email: Melissa.Hunfalvay@gmail.com

Data 640 9040

Spring 2022

Professor Steve Knode

Date: April 5th, 2022

Introduction

The dataset chosen was “car lemon dataset.csv” (Figure 1). *“Is Bad Buy”* was the target binary variable (“0” = not a bad buy, “1” = bad buy). The target variable was unevenly proportioned and skewed heavily in favor of not bad purchases (n = 64,007; 87.70%). Bad purchases totaled 8,976 of the 72,983 observations (12.30%). The imbalance percentage is 75.4%.

The purpose of the analysis was to identify the bad vehicle purchases, from good vehicle purchases. Furthermore, the dataset was used to examine models created in SAS Enterprise Miner (SAS-EM) to those in two different automated machine learning (AML) programs: Qlik AutoML and Data Robot . Specifically, data cleaning, model parameter inputs, cut-off decisions, results and evaluation of a “Champion” model are compared between the different software packages.

SAS-EM allows the user to have greater decision making control over parameters such as methods for imputation or cut-off values enabling models to be more “hand-crafted” by the analyst. AML generates models and results based on pre-set programming. A comparison between “man and machine” is therefore interesting to determine which modality creates the most sensitive model of Bad Buys.

Dataset

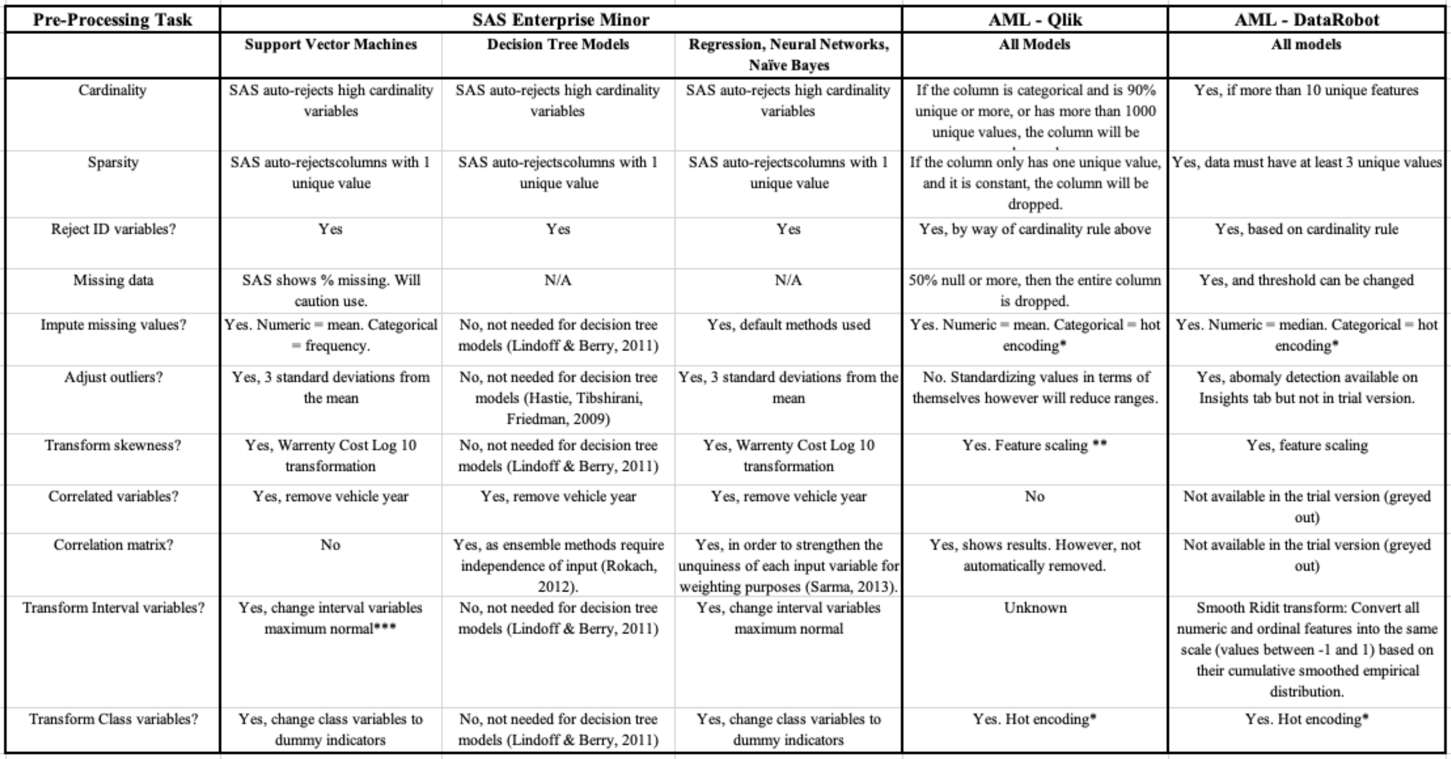
The data included various features of cars (such as color and make) and whether the cars were determined to be good or bad purchases (see Figure 1 for variable descriptions). The dataset included 34 variables and 72,983 observations. The type of variables included numeric (n = 15) and character (n = 19) variables. Of the numeric variables, 11 were interval, the other two were binary.

Missing values were present in five of the variables, ranging from 0.001% missing in the variable Transmission, to 0.02 in the MMR Acquisition variables (Figure 1). Warranty Cost has the highest skewness value (2.07; Figure 2). MMR Acquisition Auction “Clean” versus “Average” variables were highly correlated (that is above .80, Hastie, Tibshirani, Friedman, 2009, see Figure 3 and 4).

Automated Machine Learning Model Development and Discussion

A comparison of the pre-processing (data cleaning) steps for SAS-EM and AML are shown in Table 1. Input and rejected variables after cleaning are shown in Table 2 (Appendix B). Interestingly, after pre-processing, the data input to the different software programs varied considerably (e.g., number of input variables).

Table 1: Pre-Processing of Data in EM and AML[1]



Pre-Processing Task	SAS Enterprise Minor			AML - Qlik	AML - DataRobot
	Support Vector Machines	**Decision Tree Models**	**Regression, Neural Networks, Naïve Bayes**	**All Models**	**All models**
Cardinality	SAS auto-rejects high cardinality variables	SAS auto-rejects high cardinality variables	SAS auto-rejects high cardinality variables	If the column is categorical and is 90% unique or more, or has more than 1000 unique values, the column will be	Yes, if more than 10 unique features
Sparsity	SAS auto-rejectscolumns with 1 unique value	SAS auto-rejectscolumns with 1 unique value	SAS auto-rejectscolumns with 1 unique value	If the column only has one unique value, and it is constant, the column will be dropped.	Yes, data must have at least 3 unique values
Reject ID variables?	Yes	Yes	Yes	Yes, by way of cardinality rule above	Yes, based on cardinality rule
Missing data	SAS shows % missing. Will caution use.	N/A	N/A	50% null or more, then the entire column is dropped.	Yes, and threshold can be changed
Impute missing values?	Yes. Numeric = mean. Categorical = frequency.	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, default methods used	Yes. Numeric = mean. Categorical = hot encoding*	Yes. Numeric = median. Categorical = hot encoding*
Adjust outliers?	Yes, 3 standard deviations from the mean	No, not needed for decision tree models (Hastie, Tibshirani, Friedman, 2009)	Yes, 3 standard deviations from the mean	No. Standardizing values in terms of themselves however will reduce ranges.	Yes, abomaly detection available on Insights tab but not in trial version.
Transform skewness?	Yes, Warrenty Cost Log 10 transformation	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, Warrenty Cost Log 10 transformation	Yes. Feature scaling **	Yes, feature scaling
Correlated variables?	Yes, remove vehicle year	Yes, remove vehicle year	Yes, remove vehicle year	No	Not available in the trial version (greyed out)
Correlation matrix?	No	Yes, as ensemble methods require independence of input (Rokach, 2012).	Yes, in order to strengthen the unquiness of each input variable for weighting purposes (Sarma, 2013).	Yes, shows results. However, not automatically removed.	Not available in the trial version (greyed out)
Transform Interval variables?	Yes, change interval variables maximum normal***	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, change interval variables maximum normal	Unknown	Smooth Ridit transform: Convert all numeric and ordinal features into the same scale (values between -1 and 1) based on their cumulative smoothed empirical distribution.
Transform Class variables?	Yes, change class variables to dummy indicators	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, change class variables to dummy indicators	Yes. Hot encoding*	Yes. Hot encoding*

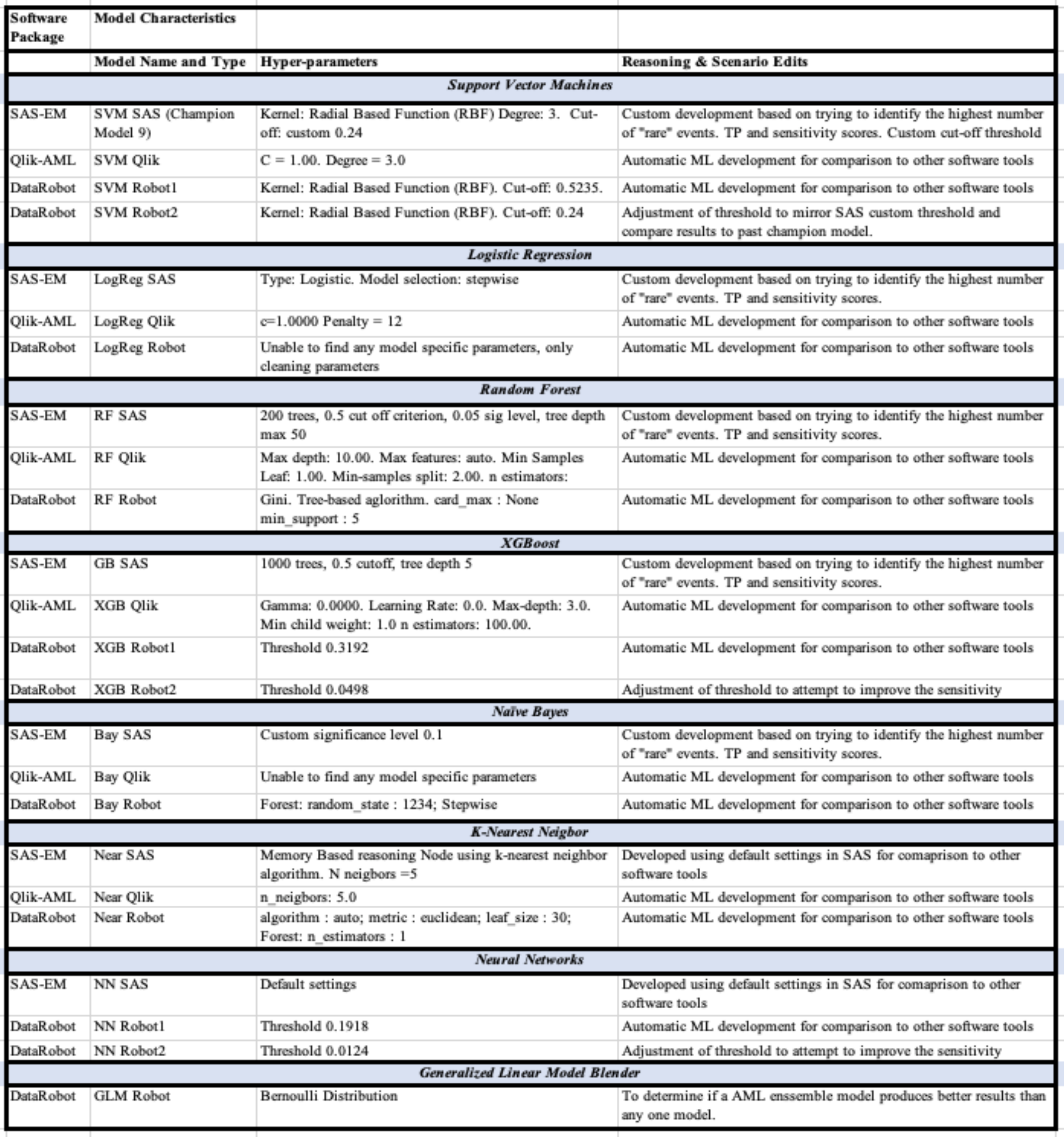
Due to limitations in the free version of the software, data was reduced prior to model generation. SAS-EM models were based on 3,648 observations (5% of the data). AML models included 4,999 observations (6.85% of the data). To effectively compare SAS-EM models to those generated by AML, the “best in class” SAS models for each model type were used (e.g. best in class regression from SAS was compared to AML regression models). Finally, a new SAS model was developed using the Memory Based Reasoning node to generate a comparative k-nearest neighbor model.

Table 3 shows the predictive models developed by SAS- EM and AML. For each SAS-EM model data was partitioned 70% training, 30% validation. For the AML models the data was partitioned into five distinct

[1] *Hot encoding pivots the categorical column into *n* number of columns, where *n* is equal to the number of unique values in the column and assigning a 1 to the appropriate column for value in each row and zero to the other columns that were generated (BigSquid: Hot Encoding)
**Feature scaling is to calculate the mean and standard deviation for each column, and then for each row calculate the number of standard deviations away from the mean that sample is (BigSquid:Feature Scaling)
*** Maximum normal is a best power transformation to maximize normality (SAS-EM: Maximum Normal)

groups called “folds” to be used for cross validation. Training data is 4 folds, each at 20%, and 20% for validation data. In AML each trained data is tested on 20% in rotation for all segments of the training data. A total of 24 models (SAS-EM: n = 7; AML: 17; Qlik: 6, DataRobot: 11, Table 2) were developed. Each program randomly sampled across the datasets. [2]

Table 2: Model Development for 24 Models Across Three Software programs (SAS-EM, Qlik, Data Robot)



Software Package	Model Characteristics		
	Model Name and Type	**Hyper-parameters**	**Reasoning & Scenario Edits**
		Support Vector Machines	
SAS-EM	SVM SAS (Champion Model 9)	Kernel: Radial Based Function (RBF) Degree: 3. Cut-off: custom 0.24	Custom development based on trying to identify the highest number of "rare" events. TP and sensitivity scores. Custom cut-off threshold
Qlik-AML	SVM Qlik	C = 1.00. Degree = 3.0	Automatic ML development for comparison to other software tools
DataRobot	SVM Robot1	Kernel: Radial Based Function (RBF). Cut-off: 0.5235.	Automatic ML development for comparison to other software tools
DataRobot	SVM Robot2	Kernel: Radial Based Function (RBF). Cut-off: 0.24	Adjustment of threshold to mirror SAS custom threshold and compare results to past champion model.
		Logistic Regression	
SAS-EM	LogReg SAS	Type: Logistic. Model selection: stepwise	Custom development based on trying to identify the highest number of "rare" events. TP and sensitivity scores.
Qlik-AML	LogReg Qlik	c=1.0000 Penalty = 12	Automatic ML development for comparison to other software tools
DataRobot	LogReg Robot	Unable to find any model specific parameters, only cleaning parameters	Automatic ML development for comparison to other software tools
		Random Forest	
SAS-EM	RF SAS	200 trees, 0.5 cut off criterion, 0.05 sig level, tree depth max 50	Custom development based on trying to identify the highest number of "rare" events. TP and sensitivity scores.
Qlik-AML	RF Qlik	Max depth: 10.00. Max features: auto. Min Samples Leaf: 1.00. Min-samples split: 2.00. n estimators:	Automatic ML development for comparison to other software tools
DataRobot	RF Robot	Gini. Tree-based aglorithm. card_max : None min_support : 5	Automatic ML development for comparison to other software tools
		XGBoost	
SAS-EM	GB SAS	1000 trees, 0.5 cutoff, tree depth 5	Custom development based on trying to identify the highest number of "rare" events. TP and sensitivity scores.
Qlik-AML	XGB Qlik	Gamma: 0.0000. Learning Rate: 0.0. Max-depth: 3.0. Min child weight: 1.0 n estimators: 100.00.	Automatic ML development for comparison to other software tools
DataRobot	XGB Robot1	Threshold 0.3192	Automatic ML development for comparison to other software tools
DataRobot	XGB Robot2	Threshold 0.0498	Adjustment of threshold to attempt to improve the sensitivity
		Naïve Bayes	
SAS-EM	Bay SAS	Custom significance level 0.1	Custom development based on trying to identify the highest number of "rare" events. TP and sensitivity scores.
Qlik-AML	Bay Qlik	Unable to find any model specific parameters	Automatic ML development for comparison to other software tools
DataRobot	Bay Robot	Forest: random_state : 1234; Stepwise	Automatic ML development for comparison to other software tools
		K-Nearest Neigbor	
SAS-EM	Near SAS	Memory Based reasoning Node using k-nearest neighbor algorithm. N neigbors =5	Developed using default settings in SAS for comaprison to other software tools
Qlik-AML	Near Qlik	n_neigbors: 5.0	Automatic ML development for comparison to other software tools
DataRobot	Near Robot	algorithm : auto; metric : euclidean; leaf_size : 30; Forest: n_estimators : 1	Automatic ML development for comparison to other software tools
		Neural Networks	
SAS-EM	NN SAS	Default settings	Developed using default settings in SAS for comaprison to other software tools
DataRobot	NN Robot1	Threshold 0.1918	Automatic ML development for comparison to other software tools
DataRobot	NN Robot2	Threshold 0.0124	Adjustment of threshold to attempt to improve the sensitivity
		Generalized Linear Model Blender	
DataRobot	GLM Robot	Bernoulli Distribution	To determine if a AML enssemble model produces better results than any one model.

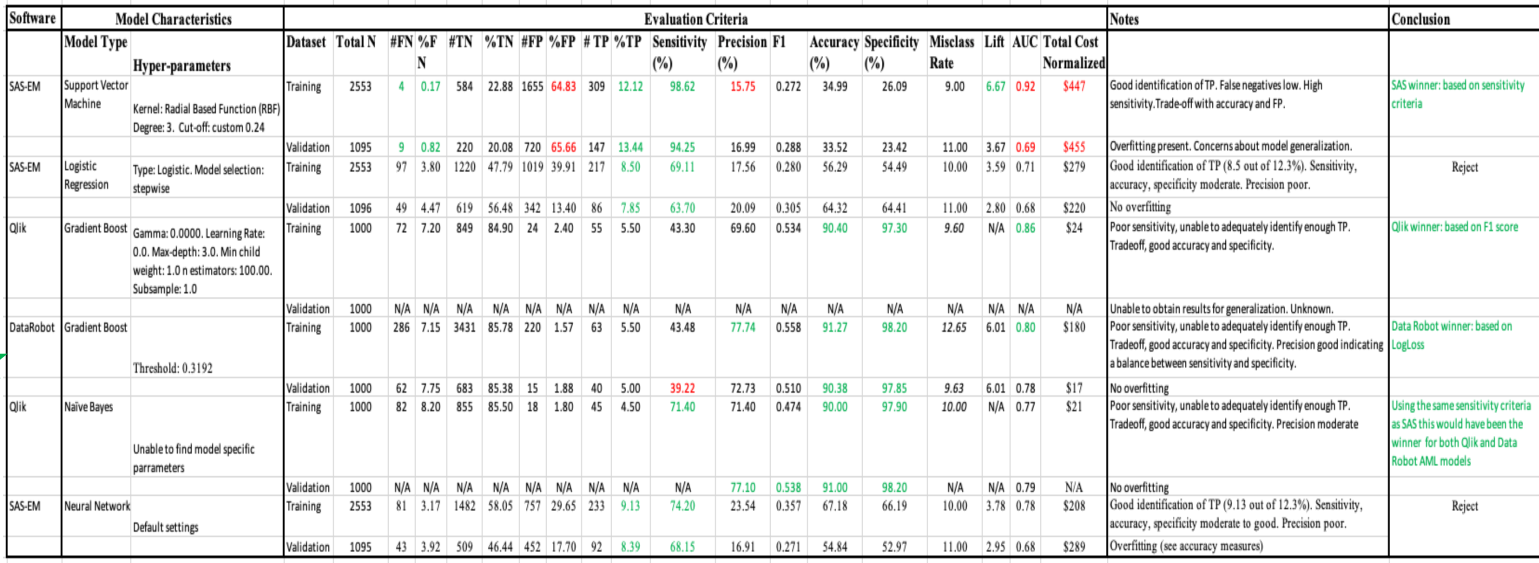
[2] Comparing models within the AML Qlik software was not able to be done as the software stopped working after the initial models were created, therefore, models between SAS, Qlik and Data Robot will be compared (rather than within Qlik).

Results

The purpose of the assessment was to identify rare (positive or "1") events, which are bad buys. Furthermore, the dataset was used to examine models created in SAS Enterprise Miner (SAS-EM) to those in two different automated machine learning (AML) programs. With these combined purposes, models were assessed using the same measures and order of importance to achieve this goal. First, true positives, second, sensitivity, third, precision, fourth, F1 score, fifth, accuracy. All these measures were examined in both training and validation data to determine the ability of the model to generalize (see Tables 3).

As all 24 models were not feasible to examine here in detail, models selected for further discussion based on the following decision-making criteria: sensitivity scores greater than 50% and/or a model selected as the winning model by the AML software and/or the prior SAS "Champion" model. The F1 score was the evaluation metric for Qlik, LogLoss (cross-entropy loss measuring the inaccuracy of predicted probabilities) is the default evaluation for Data Robot. This resulted in six models for in depth comparison (Table 3).

Table 3: Model Comparison for Top Six Selected Models



Software	Model Characteristics		Evaluation Criteria																		Notes	Conclusion
	Model Type	Hyper-parameters	Dataset	Total N	#FN	%FN	#TN	%TN	#FP	%FP	# TP	%TP	Sensitivity (%)	Precision (%)	F1	Accuracy (%)	Specificity (%)	Misclass Rate	Lift	AUC	Total Cost Normalized		
★ SAS-EM	Support Vector Machine	Kernel: Radial Based Function (RBF) Degree: 3. Cut-off: custom 0.24	Training	2553	4	0.17	584	22.88	1655	64.83	309	12.12	98.62	15.75	0.272	34.99	26.09	9.00	6.67	0.92	$447	Good identification of TP. False negatives low. High sensitivity.Trade-off with accuracy and FP.	SAS winner: based on sensitivity criteria
			Validation	1095	9	0.82	220	20.08	720	65.66	147	13.44	94.25	16.99	0.288	33.52	23.42	11.00	3.67	0.69	$455	Overfitting present. Concerns about model generalization.	
SAS-EM	Logistic Regression	Type: Logistic. Model selection: stepwise	Training	2553	97	3.80	1220	47.79	1019	39.91	217	8.50	69.11	17.56	0.280	56.29	54.49	10.00	3.59	0.71	$279	Good identification of TP (8.5 out of 12.3%). Sensitivity, accuracy, specificity moderate. Precision poor.	Reject
			Validation	1096	49	4.47	619	56.48	342	13.40	86	7.85	63.70	20.09	0.305	64.32	64.41	11.00	2.80	0.68	$220	No overfitting	
Qlik	Gradient Boost	Gamma: 0.0000. Learning Rate: 0.0. Max-depth: 3.0. Min child weight: 1.0 n estimators: 100.00. Subsample: 1.0	Training	1000	72	7.20	849	84.90	24	2.40	55	5.50	43.30	69.60	0.534	90.40	97.30	9.60	N/A	0.86	$24	Poor sensitivity, unable to adequately identify enough TP. Tradeoff, good accuracy and specificity.	Qlik winner: based on F1 score
			Validation	1000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Unable to obtain results for generalization. Unknown.	
★ DataRobot	Gradient Boost	Threshold: 0.3192	Training	1000	286	7.15	3431	85.78	220	1.57	63	5.50	43.48	77.74	0.558	91.27	98.20	12.65	6.01	0.80	$180	Poor sensitivity, unable to adequately identify enough TP. Tradeoff, good accuracy and specificity. Precision good indicating a balance between sensitivity and specificity.	Data Robot winner: based on LogLoss
			Validation	1000	62	7.75	683	85.38	15	1.88	40	5.00	39.22	72.73	0.510	90.38	97.85	9.63	6.01	0.78	$17	No overfitting	
Qlik	Naïve Bayes	Unable to find model specific parrameters	Training	1000	82	8.20	855	85.50	18	1.80	45	4.50	71.40	71.40	0.474	90.00	97.90	10.00	N/A	0.77	$21	Poor sensitivity, unable to adequately identify enough TP. Tradeoff, good accuracy and specificity. Precision moderate	Using the same sensitivity criteria as SAS this would have been the winner for both Qlik and Data Robot AML models
			Validation	1000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	77.10	0.538	91.00	98.20	N/A	N/A	0.79	N/A	No overfitting	
SAS-EM	Neural Network	Default settings	Training	2553	81	3.17	1482	58.05	757	29.65	233	9.13	74.20	23.54	0.357	67.18	66.19	10.00	3.78	0.78	$208	Good identification of TP (9.13 out of 12.3%). Sensitivity, accuracy, specificity moderate to good. Precision poor.	Reject
			Validation	1095	43	3.92	509	46.44	452	17.70	92	8.39	68.15	16.91	0.271	54.84	52.97	11.00	2.95	0.68	$289	Overfitting (see accuracy measures)	

Two finalists were considered for the overall Champion model: The SAS-SVM and Qlik- Naïve Bayes model (green starred in Table 3). Neither model is perfect. The SAS model has concerns about overfitting and cost yet has the best true positive and sensitivity rate. The Qlik model does not overfit, has moderate sensitivity, good accuracy and is the best model generally.

Receiver Operating Curves (ROC) were used in the to visually compare the tradeoff between sensitivity and specificity. An ROC index of 0.9 or greater is considered an excellent model, 0.8 to 0.9 is a good model and 0.7 to 0.8 is a fair model (Adjorlolo, 2018). Models under 0.7 are poor. The SAS training model is excellent (0.92, Figure 6) however, the validation model is poor (0.69, Figure 7). The Qlik Naïve Bayes model is fair for both training and validation datasets.[3]

Differences in pre-processing include the number of variables (SAS: n = 20, Qlik: n = 30) and the impact of correlated variables (Figures 6 & 7). The AML software does not allow the analyst to sufficiently "tweek" parameters to identify the important "rare" cases within the dataset. This is problematic and may have contributed to the reduced sensitivity for AML models in general, including the Qlik Naïve Bayes model. Variables such as 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 in SAS models. Yet it was included in the AML models.

The decision as to the Champion model comes down to the specific question being asked of the data, which is: *identify the bad vehicle purchases, from good vehicle purchases.* With this is mind, the winning model still needs to be the SAS SVM model as it does indeed identify the greatest number of bad vehicle purchases.

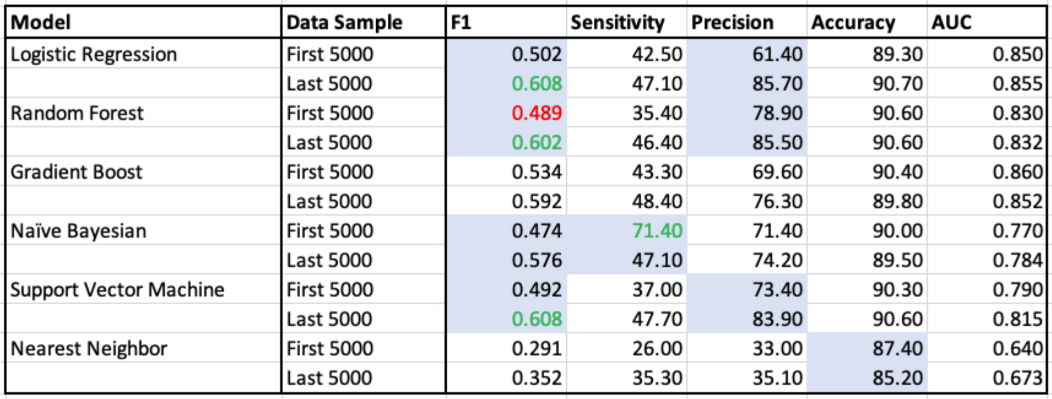
One final evaluation of the data to determine generalizability of the results and test the representative sample was undertaken using the Qlik software. A comparison of F1 scores for the first 5000 samples compared to the last 5000 samples as seen in Table 4. Results show differences in F1 scores between the different samples of data which would have resulted in different models being recommended for use by Qlik.

Overall, sample data from the last 5000 observations showed better results than the first 5000 observations. This reveals that more observations are needed (or fewer input variables) as 5000 observations still reveals enough variation in the data to result in findings that differ. When reviewing the raw data and correlation matrix (Figure 3), it was revealed that the first 5000 row analysis did not include the Transmission

[3] ROC curve visual graphs are unable to be obtained as I am unable to get back into Qlik to produce them

variable. It was rejected as all the transmission values in the observations were "auto" yet, later in the dataset (within the last 5000 rows) values were "auto" and "manual". Transmission was seen as a small but contributing factor via Worth values (Figure 4) in the SAS model and may have resulted in some differences in results between the sample data sets.

Table 4: Data Sampling Results Comparing First 5000 to Last 5000 Observations



Model	Data Sample	F1	Sensitivity	Precision	Accuracy	AUC
Logistic Regression	First 5000	0.502	42.50	61.40	89.30	0.850
	Last 5000	0.608	47.10	85.70	90.70	0.855
Random Forest	First 5000	0.489	35.40	78.90	90.60	0.830
	Last 5000	0.602	46.40	85.50	90.60	0.832
Gradient Boost	First 5000	0.534	43.30	69.60	90.40	0.860
	Last 5000	0.592	48.40	76.30	89.80	0.852
Naïve Bayesian	First 5000	0.474	71.40	71.40	90.00	0.770
	Last 5000	0.576	47.10	74.20	89.50	0.784
Support Vector Machine	First 5000	0.492	37.00	73.40	90.30	0.790
	Last 5000	0.608	47.70	83.90	90.60	0.815
Nearest Neighbor	First 5000	0.291	26.00	33.00	87.40	0.640
	Last 5000	0.352	35.30	35.10	85.20	0.673

Conclusions

In conclusion, no model was found to be a perfect solution. Even the final champion model needs to be used with caution due to possible overfitting. In advising the car sales company, I would recommend using the SAS champion model in a pilot test or "slow roll out" to see if generalizability concerns became evident. If they did, it would be recommended to include more of the data in the model to determine if overfitting was reduced.

There were some consistencies in the results across the software packages. For example, wheel type, vehicle age (or year), vehicle cost were all factors that showed important contributions to model outcomes (SAS: Figure 5, Qlik: Figure 8; Data Robot: Figure 9). These car characteristics should be considered especially noteworthy for the car sales company when considering vehicle purchases.

When evaluating the different software programs, it became especially challenging to compare models for the following reasons:

1. Input variables: both number and type of variables differed considerably (See Appendix B)
2. Cleaning parameters varied considerably between programs. SAS allowed for greater analyst control of cleaning to include the "Interactive Replacement Filter" for removal of extreme outliers in interval variables (See Figure 10). AML models varied less but did still vary (see Table 1; e.g., mean versus

median imputing of numeric variables). Changes to the AML cleaning process were fixed by the software and unable to be changed by the analyst.

3. Free version of the AML software does not allow for many changes to the hyperparameters. Where changes are permitted, it is unclear why the changes are impacting the results.
4. The model input parameters for AML software are less well understood. Even when digging into the software manuals online the depth of explanation is much less expansive than SAS documentation.
5. The different software packages evaluate the best models using varying statistics. In SAS the default is misclassification rate, in Qlik it is F1 score, and in Data Robot it is LogLoss. The evaluation criteria can be changed in SAS and Data Robot, and other evaluation criteria can be used. However, it should be noted that there is not one standard evaluation metric across packages.
6. Limitations of the Qlik software is that it does not generate Neural Network models.

Limitation in all three software packages include:

1. The amount of data (at least in the free versions) that can be used for model generation.
2. Inevitably there required some hand calculations across all packages in order to try and effectively compare models (e.g. cost, misclassification rates).

In summation, the SAS software is overall more flexible allowing the analyst to take greater control of the analytical model development but takes more time to generate models. AML software is more “hands-off” for the analyst and provides quick insights. I could see using a combination of the software packages, perhaps using the AML first to determine if models are within a “ballpark” of acceptability. If the models show promise, then transfer the data to SAS and “tweak” the models for improvement.

A final important observation was the seeming ease to which models in the AML programs could be “productionized”. Ultimately, our goal as analysts is to provide these models for real time use in the field. Therefore, an important consideration should be the ability for ease of transfer from software package to deployment.

References

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

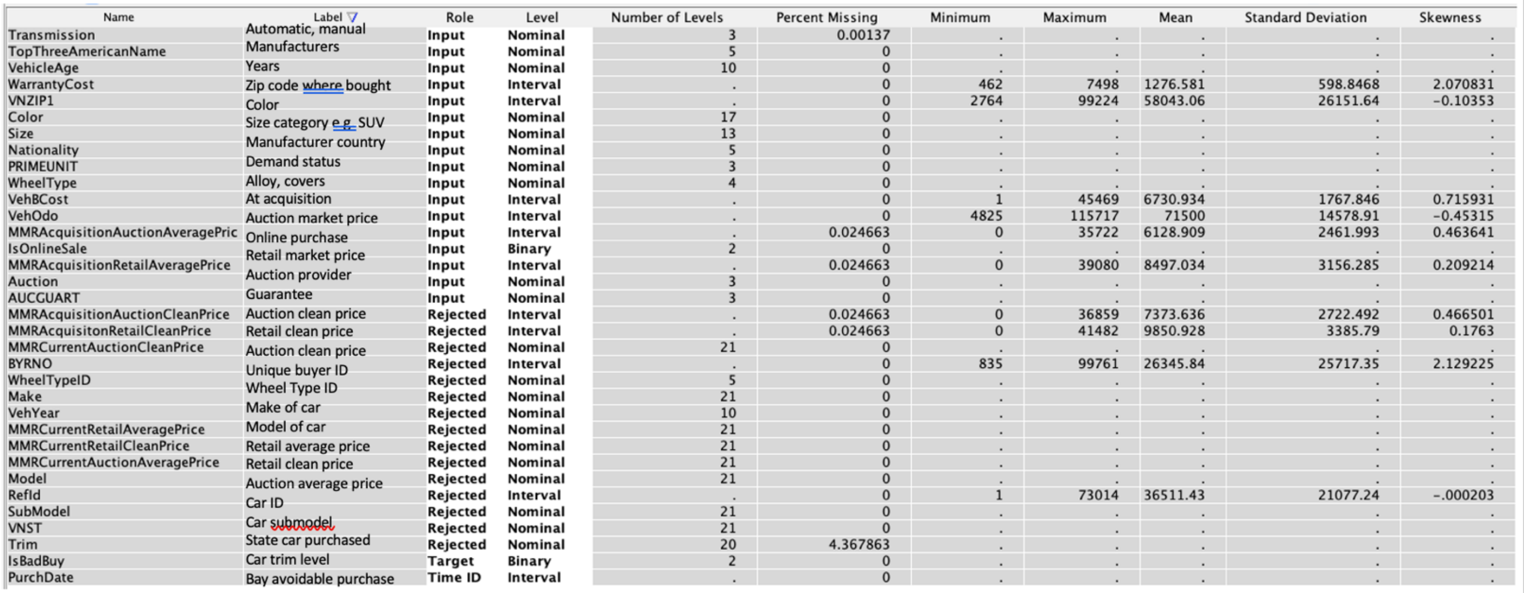
Data Robot, Inc. (2022). *UI Documentation Home.* Retrieved on April 4th, 2022, from: https://app2.datarobot.com/docs/index.html

Hastie, T.,Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, NY: New York.

Qlik, Inc. (2022). Cloud Data Integration and Analytics. Retrieved on April 4th, 2022, from: (https://kraken.bigsquid.com/).

SAS Institute Inc. (1998). *SAS Institute White Paper: Data Mining and the Case for Sampling*. Cary, NC: SAS Institute Inc. Retrieved from: https://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf

Appendix A: Figures



Name	Label	Role	Level	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness
Transmission	Automatic, manual	Input	Nominal	3	0.00137	.	.	.	.	.
TopThreeAmericanName	Manufacturers	Input	Nominal	5	0	.	.	.	.	.
VehicleAge	Years	Input	Nominal	10	0	.	.	.	.	.
WarrantyCost	Zip code where bought	Input	Interval	.	0	462	7498	1276.581	598.8468	2.070831
VNZIP1	Color	Input	Interval	.	0	2764	99224	58043.06	26151.64	-0.10353
Color	Size category e.g. SUV	Input	Nominal	17	0	.	.	.	.	.
Size	Manufacturer country	Input	Nominal	13	0	.	.	.	.	.
Nationality	Demand status	Input	Nominal	5	0	.	.	.	.	.
PRIMEUNIT	Alloy, covers	Input	Nominal	3	0	.	.	.	.	.
WheelType	At acquisition	Input	Nominal	4	0	.	.	.	.	.
VehBCost	Auction market price	Input	Interval	.	0	1	45469	6730.934	1767.846	0.715931
VehOdo	Online purchase	Input	Interval	.	0	4825	115717	71500	14578.91	-0.45315
MMRAcquisitionAuctionAveragePric	Retail market price	Input	Interval	.	0.024663	0	35722	6128.909	2461.993	0.463641
IsOnlineSale	Auction provider	Input	Binary	2	0	.	.	.	.	.
MMRAcquisitionRetailAveragePrice	Guarantee	Input	Interval	.	0.024663	0	39080	8497.034	3156.285	0.209214
Auction	Auction clean price	Input	Nominal	3	0	.	.	.	.	.
AUCGUART	Retail clean price	Input	Nominal	3	0	.	.	.	.	.
MMRAcquisitionAuctionCleanPrice	Auction clean price	Rejected	Interval	.	0.024663	0	36859	7373.636	2722.492	0.466501
MMRAcquisitonRetailCleanPrice	Unique buyer ID	Rejected	Interval	.	0.024663	0	41482	9850.928	3385.79	0.1763
MMRCurrentAuctionCleanPrice	Wheel Type ID	Rejected	Nominal	21	0	.	.	.	.	.
BYRNO	Make of car	Rejected	Interval	.	0	835	99761	26345.84	25717.35	2.129225
WheelTypeID	Model of car	Rejected	Nominal	5	0	.	.	.	.	.
Make	Retail average price	Rejected	Nominal	21	0	.	.	.	.	.
VehYear	Retail clean price	Rejected	Nominal	10	0	.	.	.	.	.
MMRCurrentRetailAveragePrice	Auction average price	Rejected	Nominal	21	0	.	.	.	.	.
MMRCurrentRetailCleanPrice	Car ID	Rejected	Nominal	21	0	.	.	.	.	.
MMRCurrentAuctionAveragePrice	Car submodel	Rejected	Nominal	21	0	.	.	.	.	.
Model	State car purchased	Rejected	Nominal	21	0	.	.	.	.	.
RefId	Car trim level	Rejected	Interval	.	0	1	73014	36511.43	21077.24	-.000203
SubModel	Bay avoidable purchase	Rejected	Nominal	21	0	.	.	.	.	.
VNST		Rejected	Nominal	21	0	.	.	.	.	.
Trim		Rejected	Nominal	20	4.367863	.	.	.	.	.
IsBadBuy		Target	Binary	2	0	.	.	.	.	.
PurchDate		Time ID	Interval	.	0	.	.	.	.	.

Figure 1: Variables in Car Lemon Dataset including a description of the variables in column called “Label”

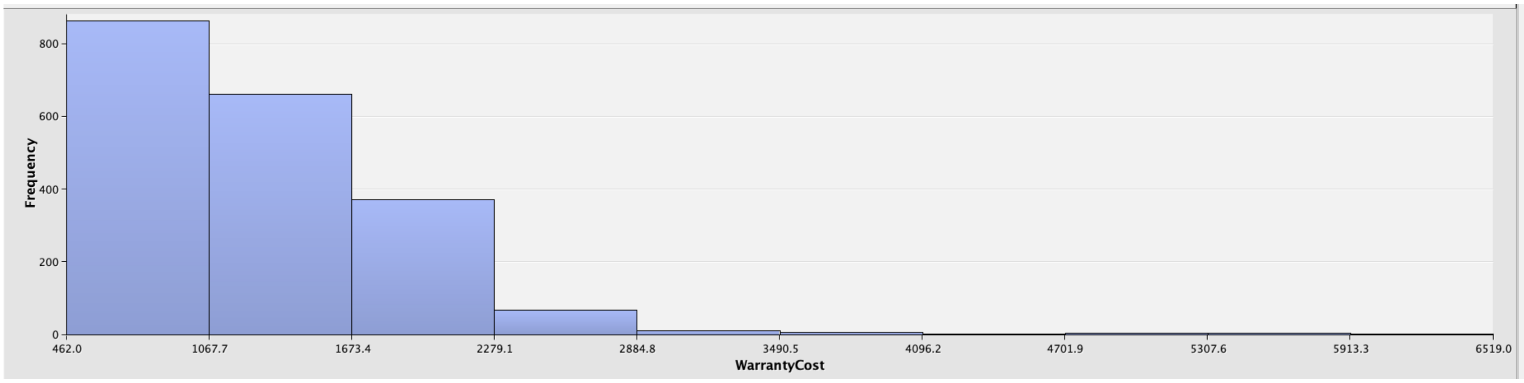


Figure 2: Skewness in the Warranty Cost variable; figure generated from SAS Enterprise Miner

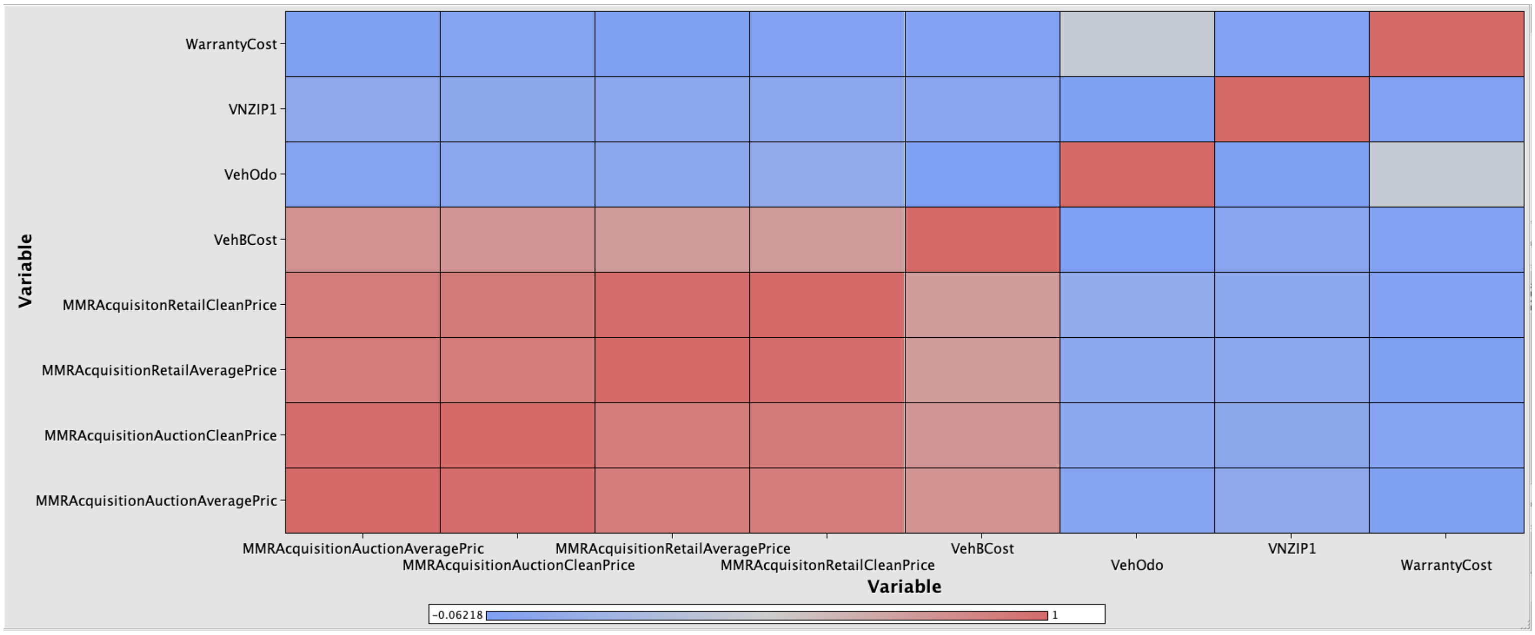


Figure 3: Correlation matrix; figure generated from SAS Enterprise Mine

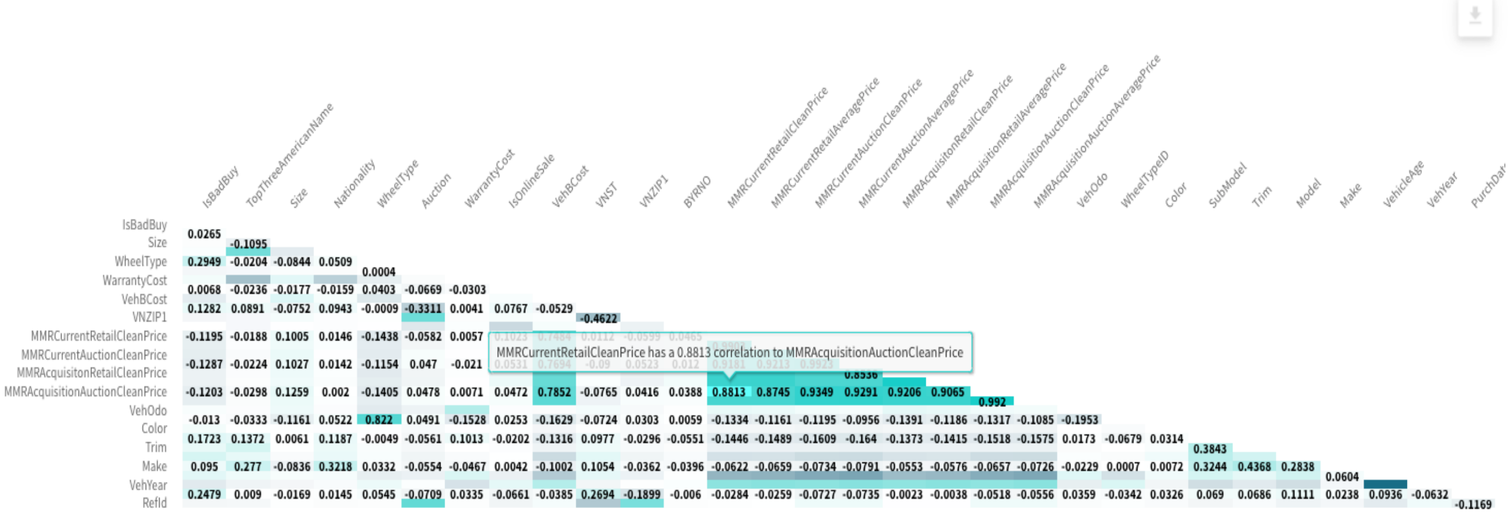


Figure 4: Correlation matrix; figure generated from Qlik. Highlighted are the most highly correlated variables.

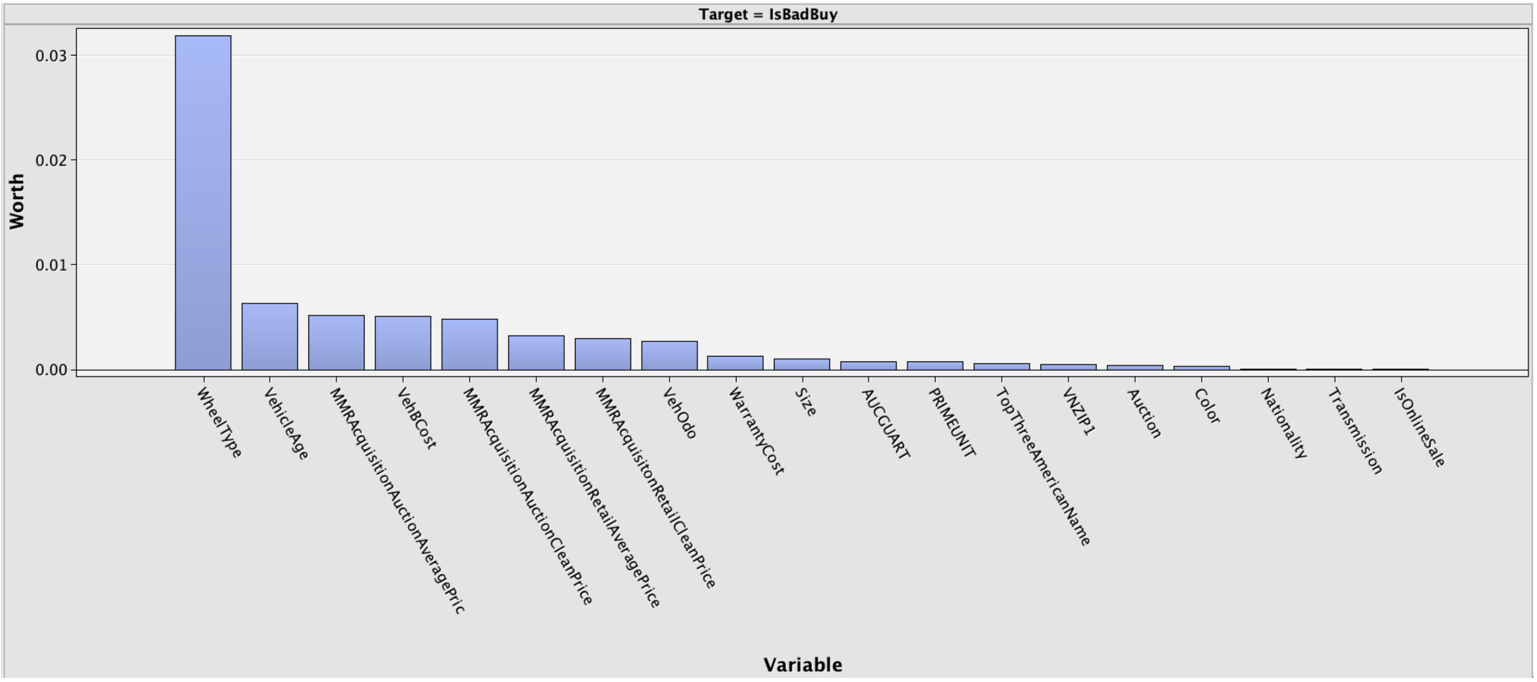


Figure 5: Worth value of the initial dataset imported via SAS.

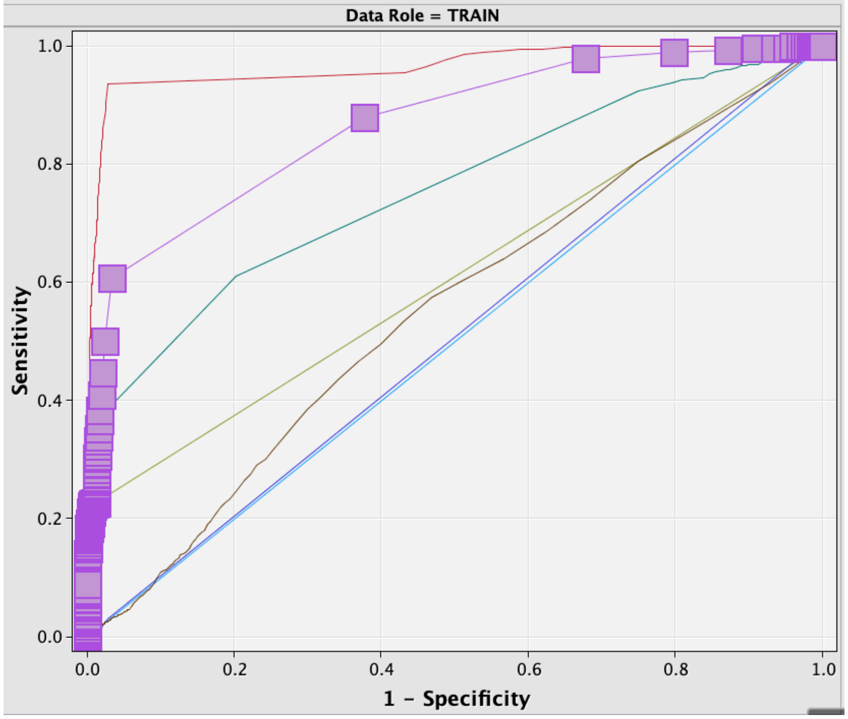


Figure 6: ROC curve for training dataset SAS SVM

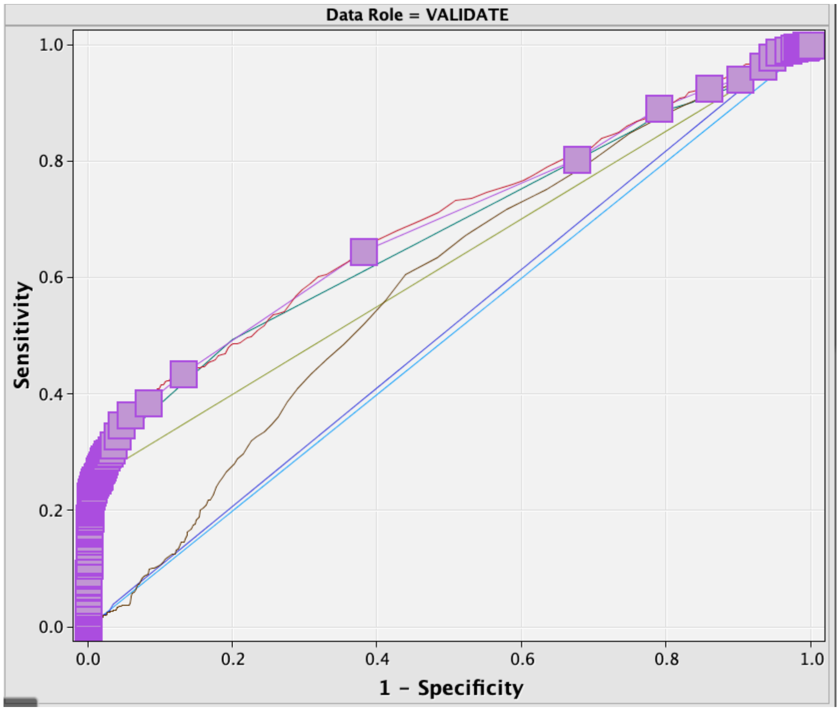


Figure 7: ROC curve for the validation training data set SAS SVM

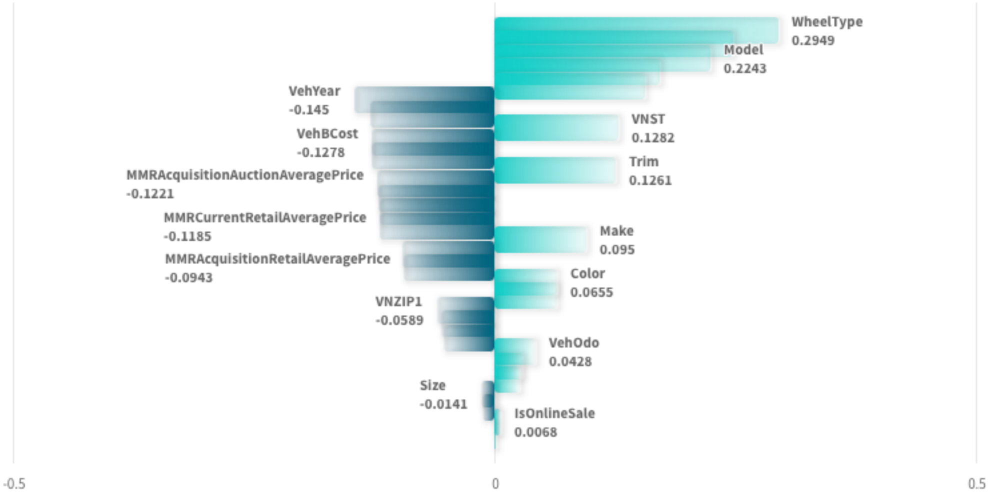


Figure 8: Target correlations showing the correlation between the target variable “Is Bad Buy” and input variables for Qlik

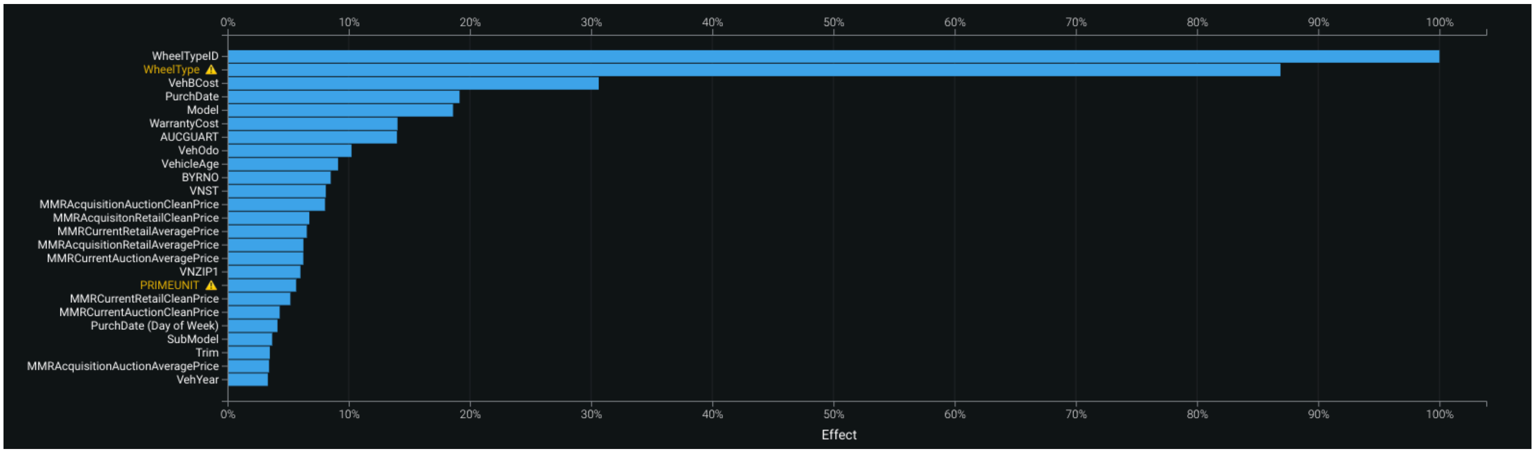
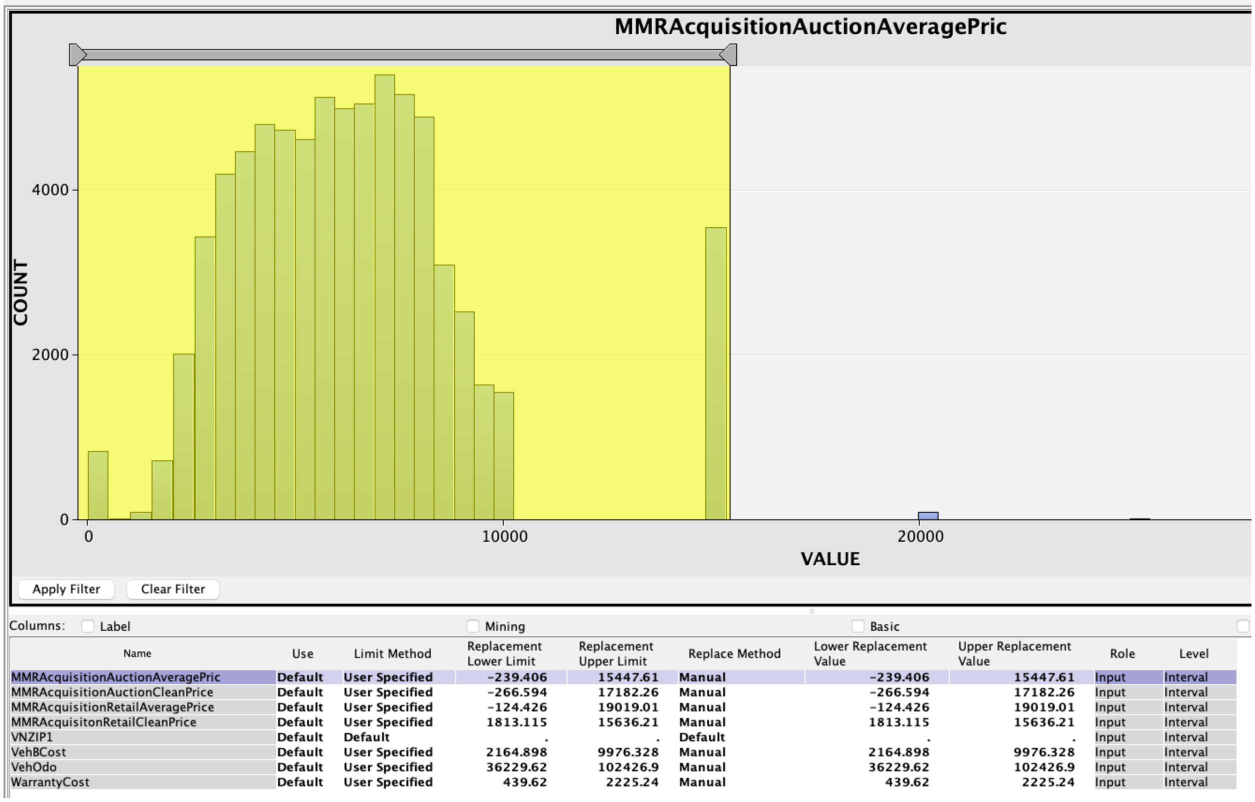


Figure 9: Target correlations showing the correlation between the target variable “Is Bad Buy” and input variables for Data Robot

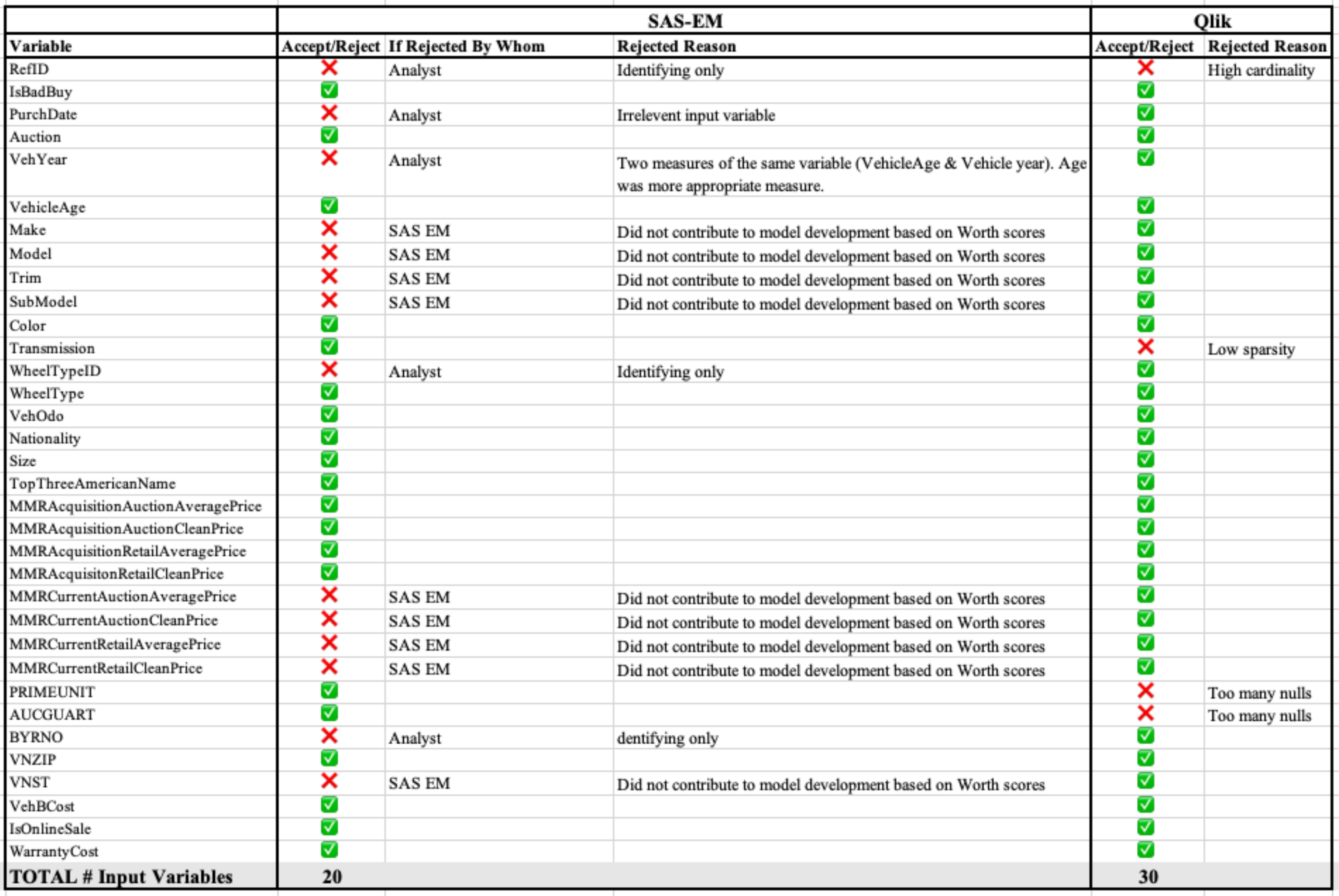


Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method	Lower Replacement Value	Upper Replacement Value	Role	Level
MMRAcquisitionAuctionAveragePric	Default	User Specified	-239.406	15447.61	Manual	-239.406	15447.61	Input	Interval
MMRAcquisitionAuctionCleanPrice	Default	User Specified	-266.594	17182.26	Manual	-266.594	17182.26	Input	Interval
MMRAcquisitionRetailAveragePrice	Default	User Specified	-124.426	19019.01	Manual	-124.426	19019.01	Input	Interval
MMRAcquisitonRetailCleanPrice	Default	User Specified	1813.115	15636.21	Manual	1813.115	15636.21	Input	Interval
VNZIP1	Default	Default	.	.	Default	.	.	Input	Interval
VehBCost	Default	User Specified	2164.898	9976.328	Manual	2164.898	9976.328	Input	Interval
VehOdo	Default	User Specified	36229.62	102426.9	Manual	36229.62	102426.9	Input	Interval
WarrantyCost	Default	User Specified	439.62	2225.24	Manual	439.62	2225.24	Input	Interval

Figure 10: EM Interactive replacement filter for removal of extreme outliers in interval variables

Appendix B: Tables

Table 2: Variable Status after Pre-Processing by SAS-Enterprise Miner and Qlik. Note SAS-EM has 20 input variables and Qlik has 30 input variables.



	SAS-EM			Qlik	
Variable	**Accept/Reject**	**If Rejected By Whom**	**Rejected Reason**	**Accept/Reject**	**Rejected Reason**
RefID	✗	Analyst	Identifying only	✗	High cardinality
IsBadBuy	✓			✓	
PurchDate	✗	Analyst	Irrelevent input variable	✓	
Auction	✓			✓	
VehYear	✗	Analyst	Two measures of the same variable (VehicleAge & Vehicle year). Age was more appropriate measure.	✓	
VehicleAge	✓			✓	
Make	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
Model	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
Trim	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
SubModel	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
Color	✓			✓	
Transmission	✓			✗	Low sparsity
WheelTypeID	✗	Analyst	Identifying only	✓	
WheelType	✓			✓	
VehOdo	✓			✓	
Nationality	✓			✓	
Size	✓			✓	
TopThreeAmericanName	✓			✓	
MMRAcquisitionAuctionAveragePrice	✓			✓	
MMRAcquisitionAuctionCleanPrice	✓			✓	
MMRAcquisitionRetailAveragePrice	✓			✓	
MMRAcquisitonRetailCleanPrice	✓			✓	
MMRCurrentAuctionAveragePrice	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
MMRCurrentAuctionCleanPrice	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
MMRCurrentRetailAveragePrice	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
MMRCurrentRetailCleanPrice	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
PRIMEUNIT	✓			✗	Too many nulls
AUCGUART	✓			✗	Too many nulls
BYRNO	✗	Analyst	dentifying only	✓	
VNZIP	✓			✓	
VNST	✗	SAS EM	Did not contribute to model development based on Worth scores	✓	
VehBCost	✓			✓	
IsOnlineSale	✓			✓	
WarrantyCost	✓			✓	
TOTAL # Input Variables	**20**			**30**	