Report:
Using Neural Networks and Boosted Neural Models to Evaluate Bad Car Purchases

Model Understanding:

A big and costly issue that auto dealers face is the risk of getting a bad used car or "kick" as they call them. Kicks could have a range of the issues with the vehicle itself or other issues such as vehicle title that could potentially be costly to the dealer. Utilizing data on used cars is important in the future in order to make future predictions on whether a car will be a "kick" or not. In order to analyze this data, this report features details concerning making and comparing neural networks and boosted neural models that will help make us of this data. In order to compare these models, ROC curve and Lift curves will be used. Confusion matrices from each of the models will be inspected and results used to evaluate the models' metrics, for sensitivity, specificity, error rate false positive and negative rates and the ROC and Lift curves. Cumulative gains curve will also be used in this evaluation.

Data Understanding:

The data set "Carvana" was provided and contained 72,983 rows. The data was screened for missing values and outliers, with 3,517 rows excluded. The data type for isBAD was changed from continuous to nominal. The data was split into three parts for validation: 60% training, 20% validation, and 20% test. In JMP, the decision tree platform was then used to re-bin several variables into Leaf Vehicle Age, Leaf MMST, Leaf VehOdo, and Leaf Vehicle Cost.

Analysis:

Upon inspection of the data and following the binning process using the decision tree platform, a model was created the following variables were chosen to be used: IsBAD, Leaf Vehicle Age, Leaf MMST, Leaf VehOdo, and Leaf Vehicle Cost.

Confusion Matrix:

Figure 1 contains the confusion matrix results with 17561 observations for neural networks and boosted neural models. Both models had the same error rate of 9%. The neural networks and boosted neural models each have a false negative rate of 9%, indicating that 9% of these predictions are false. The neural networks and boosted neural models have a has a false positive rate of 32%. This means that 32% of model predictions are false. For specificity, which denotes the percentage predicted correctly, all three of the models had a 100%. For sensitivity, all models were at 7%. Both models seem to give the same results. They have the same metrics for sensitivity, specificity, error rate, and false negative.

ROC Curve:

The ROC curves show the performance of the model. Figure 2 shows ROC curve reaching far they into the upper left corner, indicting good performance. The AUC (area under the curve) for both models very slightly different. The neural network model is represented by the red line and the boosted neural network model is represented by the blue line. For neural networks model, the AUC was 0.7570. The boosted neural models had an AUC at 0.7569, slightly lower than the neural networks model. Based on those metrics, the neural networks model is more successful at correctly classifying cases because of the higher AUC but only slightly. Even though these models are pretty similar they could be better in terms of results they produced.

Lift Curve:

Similar to the ROC curves for each model, the lift curves are also similar. Both curves exhibit the qualities of what is considering to be a "good curve" because of their left sides of each curve starting high and falling directly on the right side of the graph. Lift curves, show the effectiveness of a models prediction compared to the random 50/50 of nature of using a coin flip for prediction. As seen in Figure 3, the models is twice as accurate as simply flipping a coin for the first 40% of the data, since about 40 % of data lies above a value of 2. Figure 4 shows the measures of fit for neural networks and boosted neural models' different validation levels used in model building. Each model has a misclassification rate around 9.6% which could be better.

Cumulative Gains Curve:

Cumulative gains curves are used in modeling to determine the strengths and weaknesses of the model. In Figure 5 where isBADBuy=1, there is more deviation apart from the random 50% line in the graph. In fact, the models swing into the upper left of the graph showing a greater gain.

Conclusion:

It is important for car dealerships to use data to evaluate the risk for buying bad cars. In comparing the various models for this report, various variables were used to assess their prediction capabilities reveal some strong similarities and differences. For the confusion matrix, neural networks and boosted neural models produced the same results. For the ROC curve, all models have AUCs that are similar, with neural networks having a slightly higher AUC. For the lift curves, 40% of the data were above 2. The cumulative gains curve showed the models swinging to the far left meaning greater gains against 50/50 selection. All models predicted isBAD =1 correctly most of the time (because of high specificity), but only predicted BAD=1 correctly almost none of the time (because of low sensitivity). These models have some use but could be much better in determining whether a used car is a "bad buy".

Figure 1: Confusion Matrix Figures

Neural Network including missing value coding						
		Predicted				
	BAD	1	0	Total		
Actual	1	118	1521	1639	7%	Sensitivity
	0	55	15867	15922	100%	Specificity
Total		173	17388	17561		
		32%	9%		9%	Error
		False positive	False negative			

Neural Network Boosted						
		Predicted				
	BAD	1	0	Total		
Actual	1	114	1525	1639	7%	Sensitivity
	0	53	15869	15922	100%	Specificity
Total		167	17394	17561		
		32%	9%		9%	Error
		False positive	False negative			

Figure 2: ROC Curves

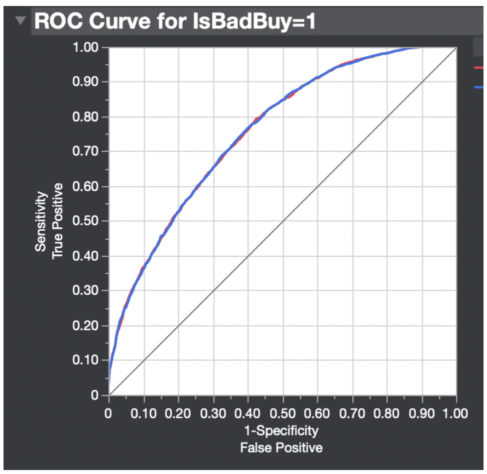

Figure 3: Lift Curves

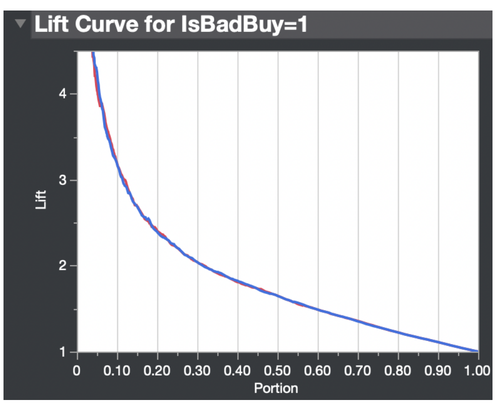

Figure 4: Measures of Fit for Validation Levels

Model Comparison Validation=Test

Predictors

Measures of Fit for IsBadBuy

Creator	.2.4.6.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
Neural Model NTanH(5) - Save Fast Formulas		0.1344	0.1751	0.282	0.2835	0.1573	0.0961	6900
Neural Model NTanH(5)NBoost(5) - Save Fast Formulas		0.1302	0.1699	0.2833	0.2841	0.1571	0.0962	6900

Figure 5: Cumulative Gains Curves