Dan Barstow

Dr. Jim Kulich

MDS 534 Final Exam

3/12/23

Prompt

You have an opportunity to offer an automobile for sale that has the following characteristics:

- 4 doors
- Medium sized luggage boot (trunk)
- Very high maintenance costs
- 4 person passenger capacity
- Medium purchase price
- Medium safety record.

You are willing to sell this auto if you believe that there is a 75% or better chance that potential customers will rate it at the acceptable level or above. Make the case for why you would or would not offer this vehicle to your customers based on the data.

Recommendations Based on Midterm

From the results of the midterm, we can say that having 4 doors means that the rating is slightly more likely to be acceptable, but not by much—same goes for the medium sized luggage boot. Having very high maintenance costs means that the car is less likely to be rated acceptable. 4-person capacity means that the car is more likely to be rated acceptable—same goes for the medium purchase price. The medium safety rating had mixed results. While the 5-Nearest Neighbor process had a classification accuracy of 79.93%, this is only a 9.93% improvement over a baseline classifier that labels everything as unacceptable. It also was only able to correctly classify 56.66% of acceptable or greater ratings, which is what we're interested in. At this point, we have a model that's better at predicting unacceptable ratings rather than acceptable ones, so we cannot be confident in its predictions. We should not put this car up for sale yet, as further analysis is required.

Rapid Miner Auto Model Results

After using the Auto Modeling feature in Rapid Miner, each model had an accuracy above 92%. Naïve bayes set a baseline accuracy of 93.3%, and gradient boosted trees achieved an accuracy of 99.4%. The ROC comparison graph was promising, as nearly every model had a nice uppercase gamma shape, Γ, indicating excellent performance against a random classifier. All models had very good recall, with all of them having 95% or higher recall. Precision was also high across models at 93% or higher. While each model is better at classifying unacceptable ratings, their performance on acceptable ratings is quite good. The AUC values for each model are very high, with the lowest being 0.984. After averaging out the importance of features from each model, safety and persons were deemed the most important features in classification, followed by purchase price and maintenance costs, then followed by luggage boot and doors. The accuracies for each model, ROC comparison graph, confusion matrix of the logistic regression, and weights for the logistic regression are shown in figures 1, 2, 3, and 4 respectively.

Looking at the modeling tabs, multiple models were either unnavigable or impossible to interpret. Deep learning and support vector machines are basically black boxes by nature. I haven't had experience with fast large margin, so I can't say much about the coefficients of that model. The random forest and gradient boosted tree models are far too complicated to begin to understand, as they consist of many smaller decision trees. But other models were easier to interpret. The decision tree was 7 layers deep, but not unnavigable, and easy to understand from a business perspective. The naïve bayes is also relatively easy to understand because it's simply a probabilistic classifier. The generalized linear model and logistic regression are easy to understand by looking at the magnitude of the coefficients, all of which made intuitive sense and agreed with the analysis from the midterm. From a business standpoint, I would prefer models like the naïve bayes, GLM, logistic regression, and decision tree due to their interpretability.

Moving to the simulator tab for each model, and inputting the profile for this problem, we find that the models were 5 to 4 in favor of an unacceptable rating. The models which do give an acceptable rating, do not do so above the 75% threshold, with the exception of gradient boosted trees. The maintenance costs were usually the deciding factor in each simulation. The performance of the logistic regression with the given profile is shown in figure 5. This would lead me to conclude that we should not have a sale for this car, in particular due to the very high maintenance costs. After changing the profile to have high rather than very high maintenance costs in the simulator tab, the models were 7 to 2 in favor of an acceptable rating, generally above the 75% threshold. The performance of the logistic regression on the suggested car profile is shown in figure 6.

Final Recommendations

My final recommendations from both the midterm and this analysis are that we should not have a sale for this type of vehicle—the very high maintenance costs will drive people away even if the other features seem like a good deal. None of the simulations on models that would classify this car as having an acceptable rating did so convincingly. However, if we to offer a comparable car with still high but slightly lower maintenance costs, then a sale would be a good idea. Another option is to allow the sale of the car with very high maintenance costs, but also offer some deal with a particular repair shop or the dealership itself to offset at least some of the maintenance costs in hopes of achieving an acceptable rating in the long run.

Figures

Accuracy

Model	Accuracy ↓	Standard Deviation
Gradient Boosted Trees	99.4%	± 0.9%
Deep Learning	97.2%	± 0.8%
Support Vector Machine	96.4%	± 0.5%
Fast Large Margin	95.5%	± 0.9%
Random Forest	95.5%	± 1.1%
Logistic Regression	94.5%	± 1.6%
Generalized Linear Model	94.3%	± 1.4%
Decision Tree	93.5%	± 1.9%
Naive Bayes	93.3%	± 0.9%

Figure 1 – The accuracy of each model.

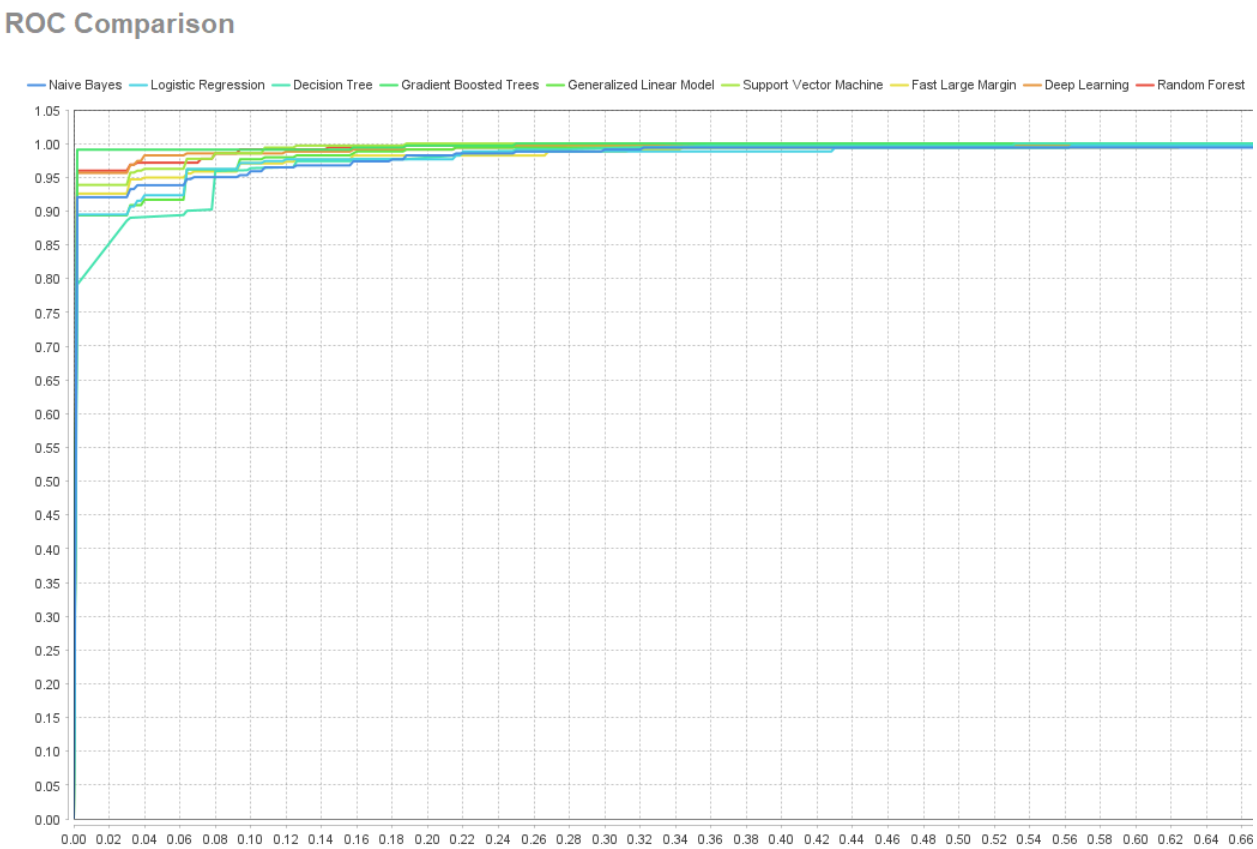

Figure 2 – The ROC comparison graph.

Logistic Regression - Performance

Profits

Profits from Model: 440 Profits for Best Option (unacc): 204 Gain: 236 Show Costs / Benefits...

Performances

Criterion	Value	Standard Deviation
Accuracy	94.5%	± 1.6%
Classification Error	5.5%	± 1.6%
AUC	98.6%	± 1.0%
Precision	95.2%	± 1.7%
Recall	97.1%	± 1.0%
F Measure	96.1%	± 1.2%
Sensitivity	97.1%	± 1.0%
Specificity	88.3%	± 3.5%

Confusion Matrix

	true acc	true unacc	class precision
pred. acc	128	10	92.75%
pred. unacc	17	339	95.22%
class recall	88.28%	97.13%	

Figure 3 – The performance of the logistic regression model after the auto modeling phase.

Logistic Regression - Weights

Attribute	Weight
Safety	0.276
Persons	0.182
Purchase Price	0.099
Maintenance Costs	0.069
Doors	0.054
Luggae Boot	0.048

Figure 4 – The weights of each feature for the logistic regression model.

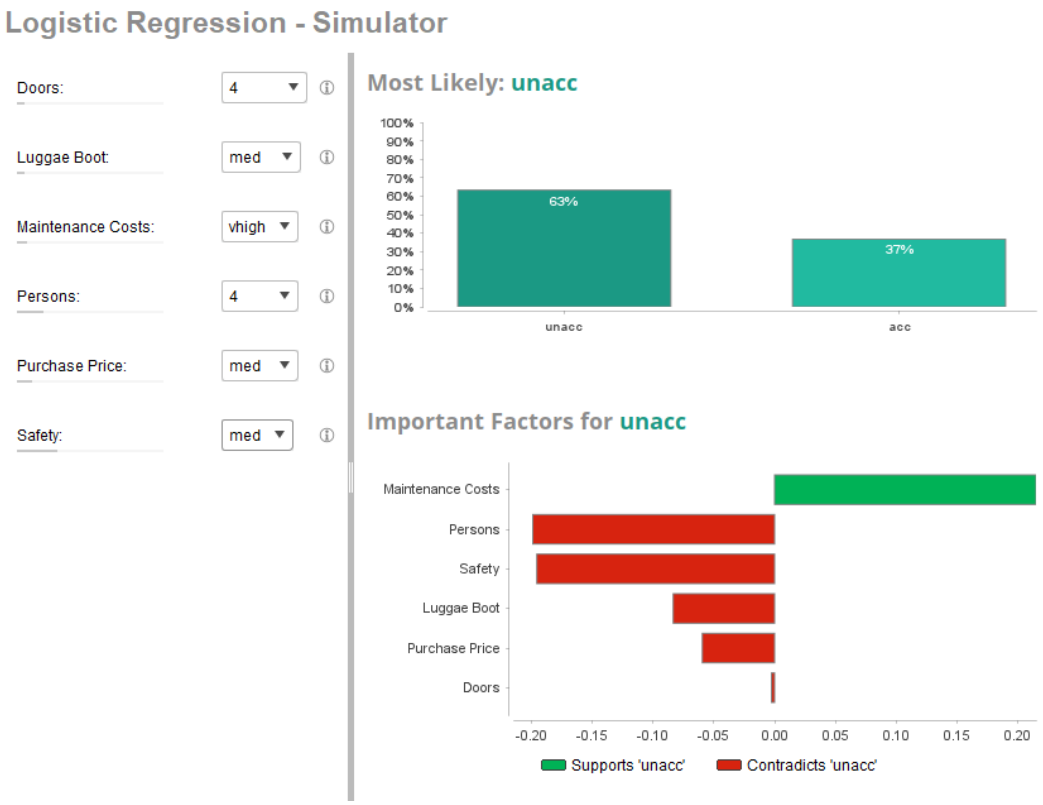

Figure 5 – The prediction for logistic regression based on the given profile.

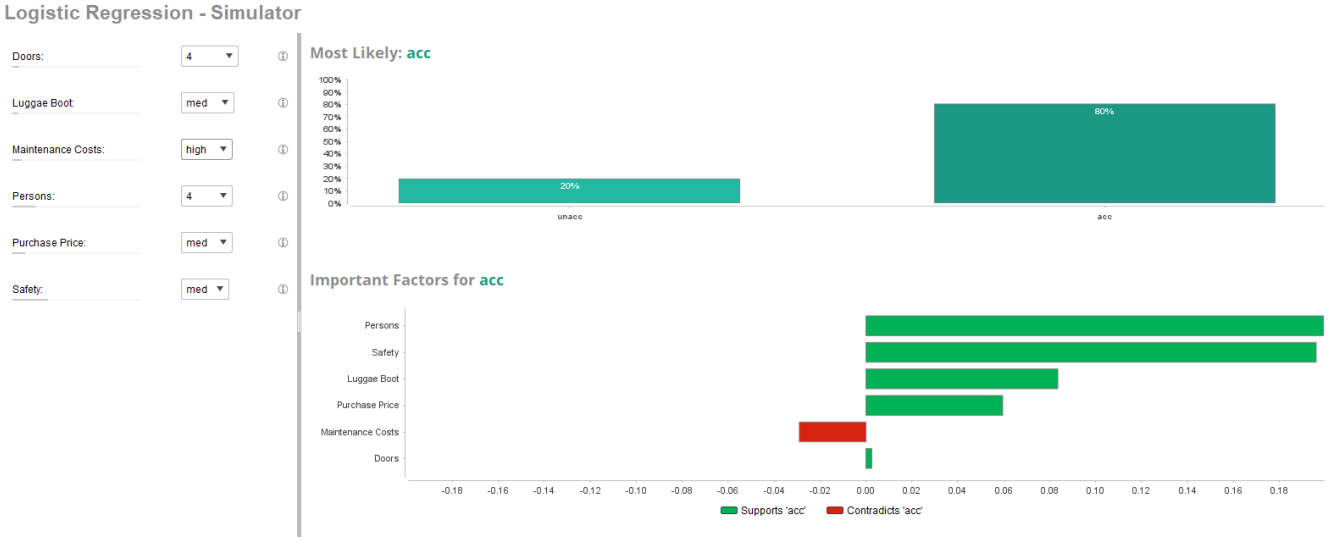

Figure 6 – The prediction for logistic regression based on the suggested profile.