Prediction of Car Prices of Federal Auctions

BUDT733- Final Project Report

Tetsuya Morito
Karen Pereira
Jung-Fu Su
Mahsa Saedirad

Executive Summary

The goal of this project is to provide buyers who attend Federal government car auctions, a simple indicator of whether or not to bid for specific vehicles, thus helping to improve their decision-making process and their chances of winning a fair bid. The data was obtained from Karl Olson, an alumnus of the Data Mining course at Robert H. Smith School of Business. The original data set contained information on 136,152 cars and included both numerical and categorical predictors. After examining the original data for feasibility, in consultation with Karl, we decided to use data on two specific car manufacturers, reducing the number of records to 12,133. The final model predictors are listed in Exhibit 1 with their descriptions.

After careful exploration of the data, we decided to transform and/or bin existing predictors to create new related predictors, to suit our problem-solving approach and modeling requirements. Transformations included *Year* and *SaleDt* to *Age in Years*, *Location* to *State*, *Miles* to *Mileage Range, and Proceeds* to *Sales Category*. This included creating bins, for *Location* based on geographical region rather than town, and for *Color* based on the base color instead of shades, in order to reduce the number of dummy variables. Bins were also created for *Miles* so that numerical ranges could be used in models such as Naïve Bayes that only accept categorical predictors. Eventually, we decided to use *Age in Years, State, Model, Miles (*or *Mileage Range), Fuel, Color* and *Sale Type* as our primary predictors to classify a specific car as belonging to the 'Average/Rough Trade-in' or to the 'Clean Trade-in or Better' *Sales Category*. These classes were created based on certain pre-determined formulae and values provided to us by Karl.

KNN, Naïve Bayes, Logistic Regression, Classification Tree, and Discriminant Analysis were the classifiers that were applied to the data set to obtain the most accurate model(s), which could potentially be used for both classification and profiling purposes. Among the above models, the Classification Tree was determined to be most accurate for classification, since it had the lowest overall error rates and was also parsimonious. Due to its comparable overall error rates and ease of interpretation, Logistic Regression was chosen as the best model for profiling, to indicate which predictors are most significant in determining the Sales Category.

Our recommendation, therefore, is to use the Classification Tree model to predict whether a new car being auctioned should be sold at the 'Average/Rough Trade-in' range or at the 'Clean Trade-in or Better' range, and to use the Logistic Regression model to understand which predictors determine the range of *Proceeds* gained when cars are auctioned.

Technical Summary

Problem/Task

The goal of this project is to help the decision making process in Federal used car auctions and indicates which variables are more important for prediction. Therefore, the task is to predict whether a car will fall into the category "Average/Rough Trade-in" or "Clean Trade-in or Better" To get the best prediction, we decide to use five classification methods: K Near Neighbors, Naïve Bayes, Logistic Regression, Classification Tree and Discriminant Analysis. Then we compare results and recommend the best performing model.

Data Analysis

The first thing we need to do is clean the data. There are more than 30 kinds of locations and colors in the data, so we create larger regions and combine sub-colors to main ones. After evaluation, only CARAVAN, IMPLALA, MALIBU and STRATUS are included in this project since their record numbers are significantly more than others. Missing data and blanks are fixed as well.

After cleaning the data, we visually analyze the data to understand which variables have different patterns with two Sales Categories, 'Average/Rough Trade-in' and 'Clean Trade-in or Better', and which variables have any relationships with them by using box charts and scatter plots. We can conclude that the age of vehicle has almost same pattern in both categories and has little predictive power in explaining differences between the two classes, while Miles show some interesting patterns (**Exhibit 3)**.

Further exploration showed that shorter mileage vehicles tend to be traded as 'Average/Rough Trade-in' range while longer mileage vehicles traded as 'Clean Trade-in or Better'. Especially 'Clean Trade-in or Better' is mainly concentrated on 50,000-110,000 miles. One of the more interesting patterns was that though we transformed Proceed to Sales Category, when we look at scatter plots of two variables, Proceed and Miles, we can see how well they separate observations between the two classes. Thus we concluded that mileage contributes to explain differences between the two classes in meaningful ways (**Exhibit 2)**.

Model Comparison

First we partition the data into 50%, 30% and 20% for training, validation and testing data set respectively. The reason to have testing data is to prevent over-fitting issues. To predict which Sales Category a specific car belongs, 'Average/Rough Trade-in' or 'Clean Trade-in or Better', and determine which variables is the most influencing to decide category, we experimented with several different types of classifiers including Logistic Regression, Discriminant Analysis, Classification

Trees, KNN and Naïve Bayes.

Then, we ran all five models using all of our variables including Age in years, State, Model, Miles (or Mileage Range), Fuel and Color after cleaning the data described above. When we run those models, we also tried several combinations of variables and pick some variables by assessing test sets error rate.

However, the results from running with 4 car models aren't very satisfying. We can see from the error rate (**Exhibit 4)** that the naïve rule is 29.31%. In Exhibit 4, even the best performing classification tree has 21.34% error rate, which has 27.19% improvement compared to naïve rule. We don't find these performances comfortable, so we decide to dig deeper into individual car model. As each car model is tried, the best improvement we have is with CARAVAN. Exhibit 4 shows the results with CARAVAN. The naïve rate for CARAVAN is 36.29% while classification tree has an error rate of 11.62%. The improvement here is 67.98%, and it gives the prediction almost 90% accuracy rates. We try all five algorithms with each car model, but only the one with CARAVAN yields better performance.

Recommendation/Conclusion

As a result, we decide to recommend classification tree (**Exhibit 5)** with CARAVAN as the model to use in the future. There are a couple of reasons we recommend classification tree. First, it has the lowest error rate among five algorithms. The error rate is only 11.62% and the improvement is 67.98%. Second, classification tree is very intuitive and straightforward; it can be easily understood and implemented in the future. Last and the most important, classification tree is a data driven algorithm. The data we acquired from Karl Olson has thousands of records, and it is only for one year. Karl has the data for the past ten years. Therefore, the predictive power of classification tree will improve as it absorbs more data and learns through it. It is for sure that Karl will get more data as time passes, so we believe classification tree is the best fit for this project.

In addition to that, Karl also expresses that he would like to know which variable is more important for prediction. For explanatory reason, we decide to run logistic regression without partitioning the data, the result is shown as **Exhibit 5.** The error rate of this non partitioning logistic regression is about 25.6%. P-value of the output shows that State_Northwest has the biggest impact among all coefficients. Furthermore, we can also see that there are many more significant variables than we see in classification tree. In the future, Karl may be able to use these variables to get a rough idea about how the auction will end up.

To sum up, our recommendation is that Karl should use classification tree as the algorithm to predict whether an auction will fall into 'Average/Rough Trade-in' or 'Clean Trade-in or Better' At the same time, significant variables found in logistic regression can be used to support judgment. We believe that this will provide the best prediction to the real life usage.

Exhibit 1: Predictors and their Descriptions

Predictor	Description	Type	New?	Included?
Year	Year in which the car was manufactured	Numerical	Original	No
Age in Years	Age of car since manufacture, in years	Numerical	New	Yes
Miles	Miles run by the car to date	Numerical	Original	Yes
Mileage Range	Binned mileage (range of 10,000 miles for each bin)	Categorical	New	Yes
Proceeds	Amount at which car was auctioned	Numerical	Original	No
SaleDt	Date on which car was auctioned	Numerical	Original	No
Cylinders	Number of cylinders in car	Numerical	Original	Yes
Location	Location of auction	Categorical	Original	No
State	Geographical region to which location belongs	Categorical	New	Yes
VIN	Vehicle Identification Number	Numerical	Original	No
Make	Manufacturer of car	Categorical	Original	Yes
Model	Name of car model	Categorical	Original	Yes
Color	Color of car	Categorical	Original	Yes
Fuel	Fuel type of car (Ethanol – ETH or Gas – GAS)	Categorical	Original	Yes
Sale Type	Live auction or auction over the Internet	Categorical	Original	Yes
Sales Category	'Average/Rough Trade-in' or 'Clean Trade-in or Better'	Categorical	New	Yes

Exhibit 2: Relationships between predictors, and between predictors and Proceeds

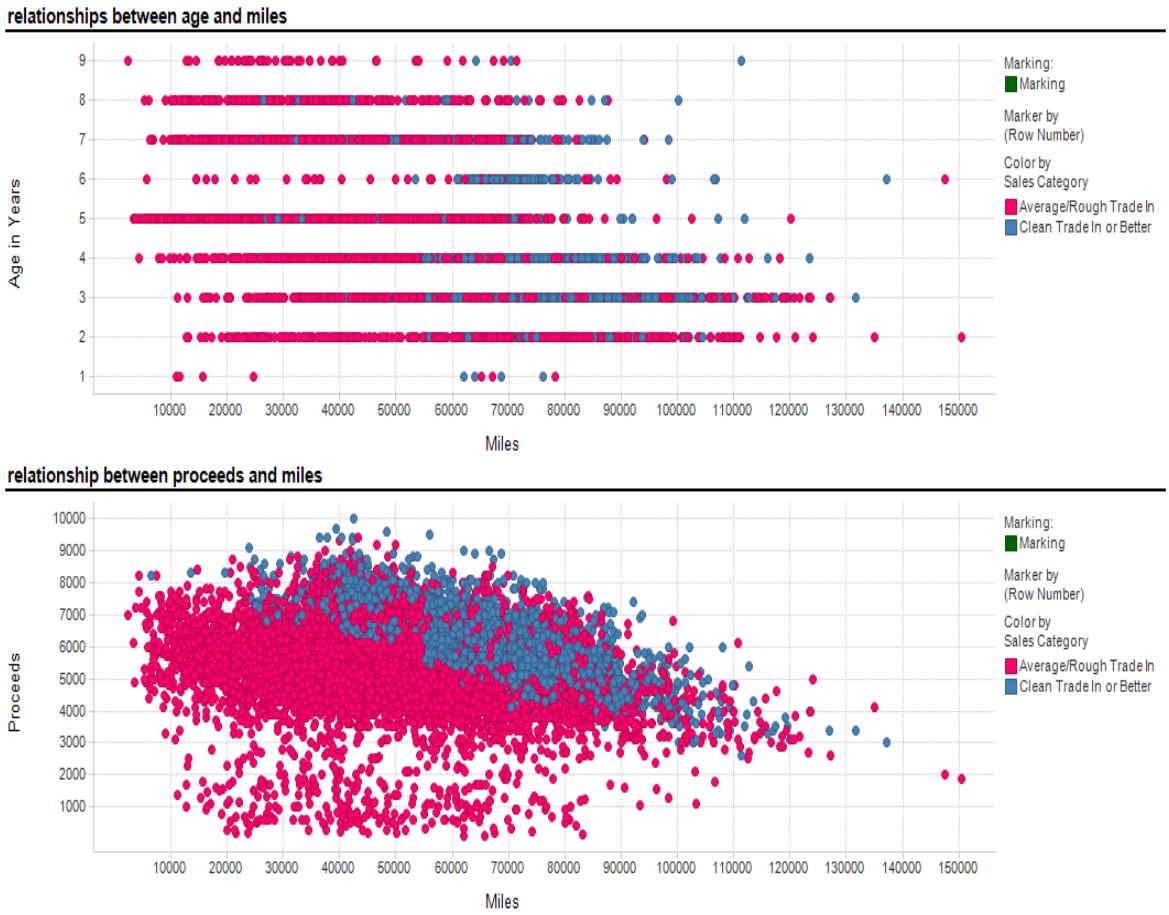
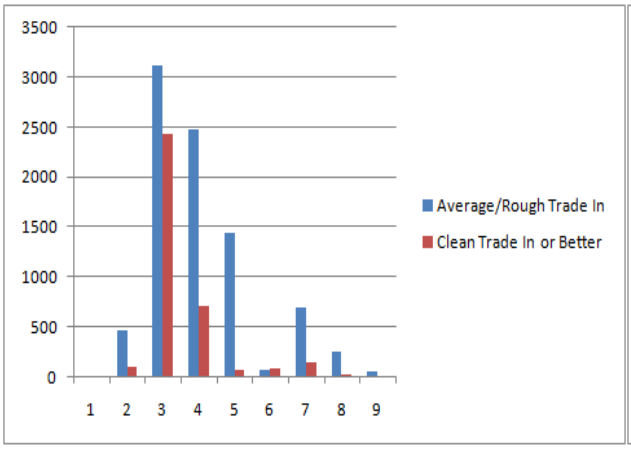


Exhibit 3: Distribution of prediction classes

Age in years



Miles

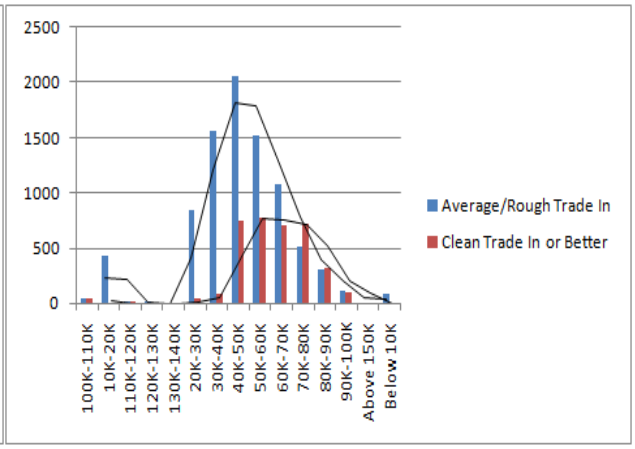
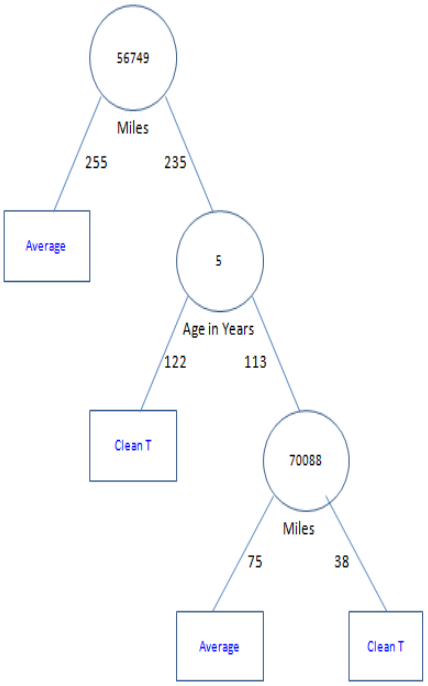


Exhibit 4: Model Comparison

Error Rate	4 car models	Improvement	Caravan	Improvement
Naïve Rule	29.31%	NA	36.29%	NA
CT	21.34%	27.19%	11.62%	67.98%
LR	21.84%	25.49%	13.46%	62.91%
KNN	23.32%	20.44%	16.82%	53.65%
NB	24.02%	18.05%	13.46%	62.91%
DA	25.71%	12.28%	13.46%	62.91%

Exhibit 5: Classification Tree and Logistic Regression



Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.14360505	0.18952082	0.44861442	*
State_Mid-West	-0.73529267	0.08010637	0	0.47936514
State_North-East	0.85433447	0.12420381	0	2.34980989
State_North-West	2.2328167	0.53961253	0.00003506	9.32609749
State_South	-0.55373871	0.09664375	0.00000001	0.5747968
State_South-East	-0.59091103	0.07208678	0	0.55382252
State_South-West	0.30048233	0.09211288	0.00110586	1.35051
Age in Years	0.10373389	0.02297814	0.00000635	1.10930526
Miles	-0.000042	0.00000153	0	0.99995798
Color_GN	-0.390802	0.07527568	0.00000021	0.67651409
Color_WH	0.20398466	0.06595763	0.0019837	1.22627938
Cylinders	0.58939332	0.02958546	0	1.80289435