Prediction if a car is a Kick!

Onkar Daiv (odaiv@nyit.edu)
Priyanka Jadhav (prijadhav@nyit.edu)

Dept. of Computer Science,
New York Institute of Technology,
Manhattan, NY

Abstract *– This paper aims at analyzing the car auction database by building a model to predict if the car purchased at the auction is 'good buy' or not. Various classification algorithms like Random Forest, J48, Naïve Bayes etc. have been compared to find out which is the more accurate classifier that predicts the car which is in best condition considering the strong attributes. Results suggests that Random forest classifies most accurately with a low mean absolute error on the test data.*

Keywords—kicks; Random Forest; J48; Chi-squared Ranking; SMOTE; Decision Tree; oversampling;

1. Introduction

Dealers often tend to purchase cars at an auction that are bad buys termed as 'kicks' [1]. We carry out a predictive analysis for the car auction database to determine if the car purchased is a good buy. The results of this analysis to predict if the car is a good buy or not will enable the dealers to purchase relatively good cars at an appropriate price thereby helping them to sell it to the customers with the best selection possible. We build our classification model based on attributes
such as "MMRCurrentRetailAveragePrice", "MMR AcquisitionAuctionAveragePrice", vehicle year, VNST, purchasedate, warranty cost and wheeltype. We used attribute ranking method to find the topmost attributes with highest information gain. The raw database was obtained from kaggle.com consisting of 32 attributes and 72984 instances on which performed major data cleaning including eliminating null values, data rollup [2] modifying some attributes relevant to the analysis.

Identifying attributes that are most important to our analysis is an important task. With the help of predictive analysis, we have determined the effect of various attributes were useful in our hypotheses. We have implemented an application that makes use of the classifiers such as Random Forest, J48 etc. to predict whether the car is a good buy or a bad buy. Results show which attributes influence the decision whether the car is to a 'good buy'. Our approach follows various steps that include – Data acquisition, Data cleaning and preparation, identifying relevant attributes, building predictive model[3] and analyzing results.

2. Hypotheses

The major hypotheses that serves as a baseline for this project is – "**To Predict whether a car at an auction is a good buy or not**". By this hypotheses, our aim is to give the best estimation to the dealers at a car auction so that they can purchase cars that are in the best shape. After selecting the top ranked attributes conducive to our hypotheses, our objective is to build an appropriate classifier that gives the best classification of the cars - whether they are a good buy or not and that gives decently high prediction rates.

3. Data Preparation

The car auction database was obtained from kaggle.com.

The complete link for database is– https://www.kaggle.com/c/DontGetKicked/data

The dataset initially comprised of 32 attributes and 72984 instances which were reduced to 18 attributes after performing data cleaning. We split the data into 70% training set, 15% testing set and remaining 15% validation set. Major portion of data cleaning and preparation included –

1. The "PurchaseDate" attribute was turned categorical into 12 months specifically to examine the prediction results throughout the year.
2. Null records were deleted for attribute prime unit whereas null values for attributes wheeltypeid, transmission, nationality, size and color were replaced by most commonly occurring values.
3. In the Make Attribute, Toyota and Toyota Scion values were merged.
4. The AUCGUART attribute contained three categorical data i.e. Red, Green and Null. So we had assumed the null to be Yellow but later realized that it consisted of more than 90% of the data and wouldn't be helpful in the analysis.
5. The null values in attributes MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice,MMRAcquisiti onRetailAveragePrice,MMRAcquisitonRetailCleanPr ice,MMRCurrentAuctionAveragePrice,MMRCurren AuctionCleanPrice,MMRCurrentRetailAveragePrice and MMRCurrentRetailCleanPrice were replaced by average values.

4. Data Mining

4.1 Data Mining Techniques

Data Mining is the process of exploring large amount of data in search of consistent patterns and finding relationships among attributes and applying the detected patterns to new sets of data. In this project, the data mining technique used is classification where Random Forest, J48, Decision tree and adaBoostM1 were compared to classify the cars if they were good buys or not. Using the attribute ranking approach [4], we used the combination of Chi-squared Ranking Filter, Information Gain Ranking Filter [5] and Gain Ratio feature evaluator to come up with the most valuable attributes.

4.2 Data Mining Tools –

We used the following data mining tools –

1. Weka
2. R

5. Results

Initially the training dataset consisted of more than 85% of the instances that classified the output as 'Good buy'. Due to this imbalance in the dataset, oversampling using the SMOTE filter in weka was applied that increased the instance of car being a bad buy from 10% to 35% with the enhanced dataset. Although performing undersampling would not be a good choice as we would require to drop the important chunk of instances classified as YES(IsGoodBuy) to match with the instances with prediction as NO.

We applied 5 different classifiers of the most highly ranked attributes and analyzed the results. After comparing the results for oversampled, undersampled and normal data we obtained the following statistics –

	ACCURACY			MEAN ABSOLUTE ERROR			ROC AREA		
	Under Sampling	No Sampling	Over Sampling	Under Sampling	No Sampling	Over Sampling	Under Sampling	No Sampling	Over Sampling
NAIVE BAYES	60.53	76.02	67.57	0.404	0.258	0.335	0.650	0.656	0.726
J48	61.564	88.0193	86.904	0.460	0.214	0.238	0.631	0.5	0.844
RANDOM FOREST	56.926	87.50	86.904	0.4353	0.2023	0.2275	0.572	0.659	0.912
ADA BOOST	61.647	88.01	73.93	0.463	0.206	0.36	0.655	0.667	0.783

Table 5.1 Results comparison table

	ACCURACY		
	Under Sampling	No Sampling	Over Sampling
J48	51.564	77.053	86.904
RANDOM FOREST	49.946	84.150	82.502

Table 5.2 Second Analysis

From the above comparison, it appears that J48 decision tree is the most accurate classifier of all but this was majorly due to the three strong attributes in the dataset that is "wheel_type" and "Vehicle_Age"/year. Hence, the analysis was again done by removing these two variables and checking for the consistency in accuracy between Random forest and Decision tree and the results suggested that there was a significant drop in the accuracy of decision tree but accuracy of random forest remained close to the previously predicted.

Analysis and plotting of Important Variable:

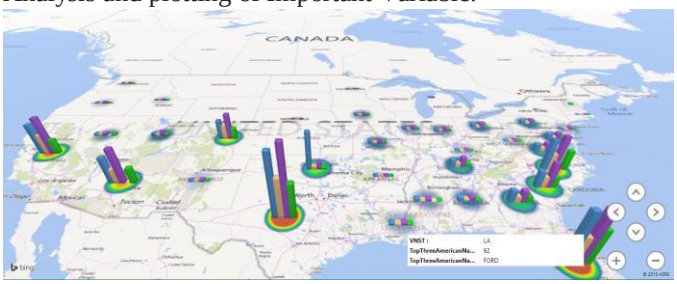

Fig5.1 Plot of Class Variable with top 3 car manufacturers

The figure above helps us to find out the top car manufacturers that we must buy the car from which are good buys. Also, we can analyse that which states in united states are having the maximum amount of cars sold in auction and how many of them turn out to be good buys.

6. Lessons Learned

From this analysis we learned that predictive modelling is an important aspect of any data mining project where determining most important attributes is the key. Similarly, transforming raw data into data that is imperative to our hypotheses by the means of data cleaning and preparation forms the basis of any data mining project. The major problem that we encountered was the 'Yes' values biasing the database for which we used over-sampling and under-sampling to balance the dataset. To conclude, by further enhancing the predictive performance, the decision of predicting if the car purchased at an auction is a good buy or not may prove significant in helping the dealers and ultimately the customers to purchase cars in the best possible condition.

7. References

[1] Albert Ho, Robert Romano, Xin Alice Wu, Don’t Get Kicked - Machine Learning Predictions for Car Buying.

[2] Paresh Patel, Shlomish Consulting, Inc., Wauconda, IL, Custom Rollup: When the Cube’s Default Behavior Doesn’t Do the Right Job!,Paper 027-2012

[3] Finlay, Steven (2014). *Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods* (1st ed.). Basingstoke: Palgrave Macmillan. p. 237.ISBN 1137379278.

[4] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit , LineUp: Visual Analysis of Multi-Attribute Rankings.

[5] Mark A Hall and Geoffrey Holmes, Bench Marking Attribute Selection Techniques for Data Mining, July 2000