[Don't Get Kicked!](https://www.kaggle.com/c/DontGetKicked)

# (Kaggle Project by www.carvana.com)

# Team Members

Mayank Jaglan [mjaglan@umail.iu.edu](mailto:mjaglan@umail.iu.edu)

Raghava Reddy [ramarvab@umail.iu.edu](mailto:ramarvab@umail.iu.edu)

Sachin Sable [ssable@umail.iu.edu](mailto:ssable@umail.iu.edu)

## Faculty Advisor

## Professor Predrag Radivojac predrag@indiana.edu

# Objectives & Significance:

The main objective of this project is to predict if the car purchased at the Auction is a Kick (bad buy). This help the dealers to choose the best cars in the auction which in turn will maximize the profits and minimize the risk. Next, in quest of creating an optimal mining model, we will do attempt a number of popular classification approaches with test and trial based tweaks for the sole purpose of improving the accuracy and/ or computational performance.

Why is it important

One of the biggest challenges of an auto dealership purchasing a used car at an auto auction is the risk of that the vehicle might have serious issues that prevent it from being sold to customers. The auto community calls these unfortunate purchases "kicks".

Kicked cars often result when there are tampered odometers, mechanical issues the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. Kick cars can be very costly to dealers after transportation cost, throw-away repair work, and market losses in reselling the vehicle.

Modelers who can figure out which cars have a higher risk of being kick can provide real value to dealerships trying to provide the best inventory selection possible to their customers.

Motivation for doing it

For the above reasons, car dealerships can benefit greatly from the predictive powers of Data Mining. If there is a way to determine if a car would be kicked a priori, car dealerships can not only save themselves money, but also provide their customers with the best inventory selection possible. Providing a model with very high accuracy will definitely help the dealers to understand the trends of customer and yields maximum profit.

# Background:

Automotive manufacturing and reusing requires a fundamental understanding of the market, the trends, the moods, and the changing consumer tastes and preferences are fundamental to competitive. The automotive industry is growing rapidly. Globalization continues to introduce new opportunities and increase competition. Dealers want better and more extended products range to sell to their customers. But one of the biggest challenge dealers are facing is to estimate the price for a used car. This estimated price depends on several factors like

* Family cars are better sellers than car for single people like sports cars.
* Some car colors sell much better than others. For example, green is the poorest seller and white the best.

One of the biggest challenges of an auto dealership purchasing a used car at an auto auction is the risk of that the vehicle might have serious issues that prevent it from being sold to customers. The auto community calls these unfortunate purchases "kicks".

Kicked cars often result when there are tampered odometers, mechanical issues the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. Kick cars can be very costly to dealers after transportation cost, throw-away repair work, and market losses in reselling the vehicle.

Modelers who can figure out which cars have a higher risk of being kick can provide real value to dealerships trying to provide the best inventory selection possible to their customers. If the probability of purchasing a kicked car can be reduced, then dealers would be able to provide better inventory for customers, minimize their costs, which will in turn increase their profits.

To understand the significance of this problem, consider the below example.

|  |  |  |
| --- | --- | --- |
| Parameter | Amount | Description |
| Number of cars bought and sold by auto dealers | 20,000 | Assumption |
| Percentage of kicked cars | 12.30% | From the dataset |
| Number of kicked cars | 2460 | Number of cars which are bad buys |
| Average price per car sold | $10000 | Assumption |
| Profit if not a bad buy | $2000 | Assumption |
| Profit if it’s a bad buy | $500 | Profit drops 1500$ in case of bad buy |
| Loss of potential profit | $1500 |  |
| Total loss of profit | $3,690,000 |  |

So, the dealers are losing $3,690,000 for 20,000 cars bought in the auction. That’s very huge amount for the dealers. If we can increase the accuracy of the random classifier, then there will be a huge profits for the dealers.

For example, if we increase the accuracy of classifier, then below are the profits obtained by the dealer for every 20,000 cars bought in the auction

|  |  |
| --- | --- |
| Accuracy | Profit gained |
| 1% | $300,000 |
| 2% | $600,000 |
| 3% | $900,000 |
| 4% | $1,200,000 |

If the number of cars bought in auction is more, then there will be huge increase in profits for the dealer.

# 5.Methods:

### Describe your data and how you obtained it?

We are given an anonymized dataset containing categorical, ordinal and numeric variables available when the used cars were purchase by the users. All string type variables are either categorical or ordinal variables. There are total of 32 features and 72983 train set records. The "IsBadBuy" column in the train set is the variable to predict. It is equal to 1 for pre-owned cars with faults.

We categorized the variables into 3 selection types

1. Car specifications
2. Auction specific
3. Price related features.

#### Data Integration/Data Reduction:

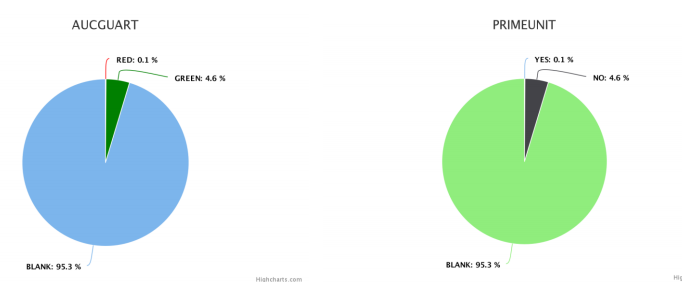
1. **Poor Quality Variables:**

**PRIME UNIT**

From the above summary of the data, we found that feature “PRIME UNIT” which identifies whether the vehicle has higher demand than a standard purchase has most of the records as NULL and only few rows have data “Yes” or “no”.

**PURCH DATE:**

44939 records in the dataset have “purchdate” which tells the date the vehicle was purchased at the auction as NULL i.e. 60% of the records. Also we believe purchase date will not provide any value to the predicitons.



1. **Redundant data :**

“VehYear” and “VehicleAge are two features with the same meaning which tells the date the vehicle was bought. So, we excluded “VehYear for our analysis.

Also the features “VNZIP” and “VNST” are similar, so we removed VNZIP from the features.

#### Data Cleaning:

For rest of the features, we classified them into two types

1. Numerical
2. Nominal or categorical

We have written the code for removing the null values from the data. All the numerical features which contain null values was replaced with mean value of that feature and all the nominal or categorical features were replaced with mode.

Please check the code preprocess.py which was written to implement the above functionality

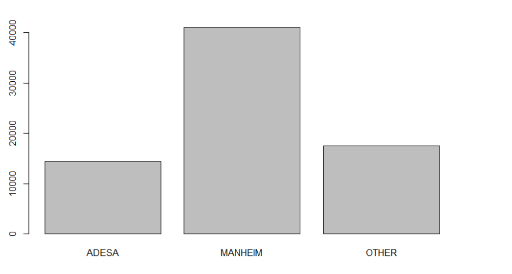
##### Data Transformation:

**Transmission Feature:** Transmission Feature have 2 Categories AUTO and MANUAL, but some of the records in the dataset for this feature have lower case variables Auto and Manual. So we converted them into upper case categories AUTO and MANUAL.

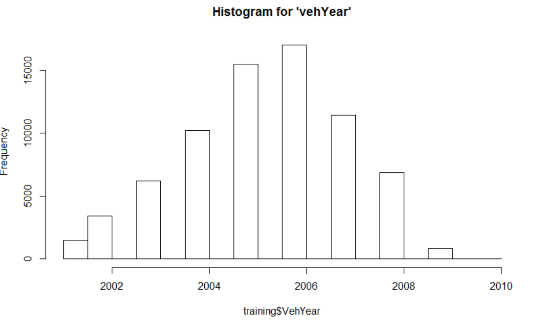
**AucGUart Feature** : It should have 3 categories Green, Yellow and Red. But in the given dataset there were Red, Green and Yellow. But 95% of the data is yellow which will not help for the model. So we ignored this feature as well.

##### Feature Selection:

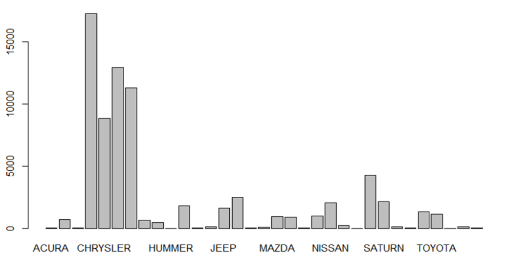
* Auction has 3 distinct value which are decently spread.



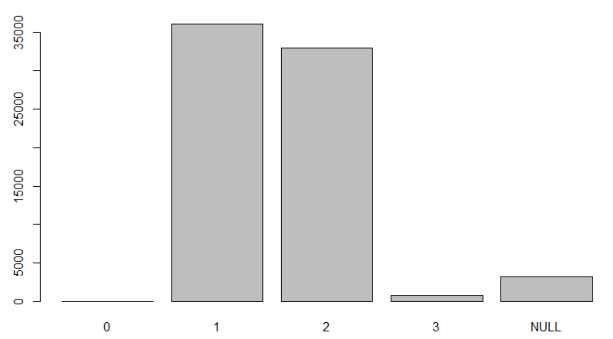
* VehYear, tells us information about the purchase of the vehicle and it follows Gaussian distribution.



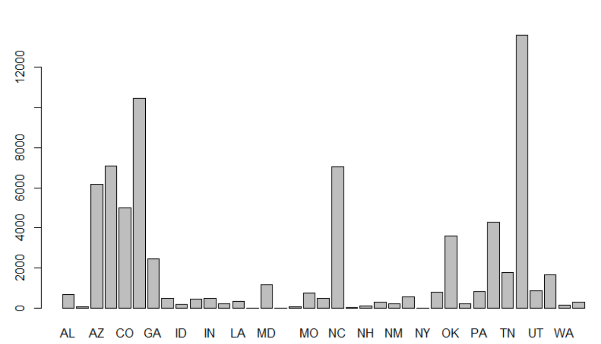
* Make is left skewed and manufacturer would have impact on the resale price



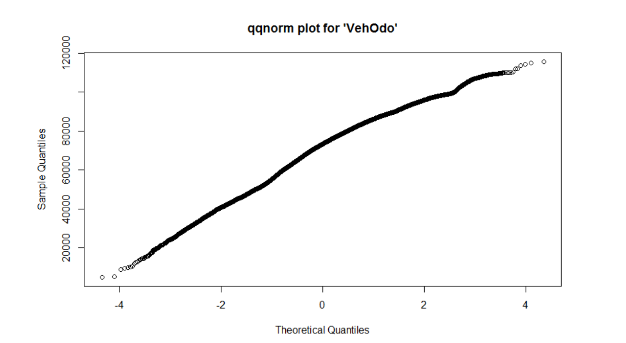
* Color is useful because many people have preferences on the desired color
* Wheel type ID have similar type distribution between two values.



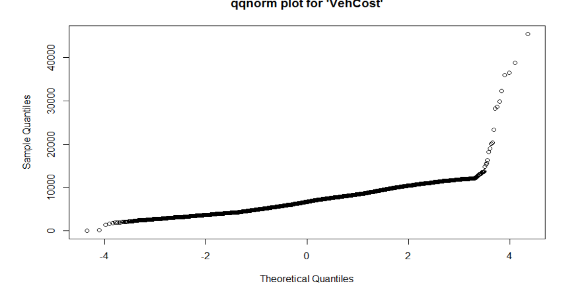
* Nationality indicates that people from certain countries will not maintain their vehicles properly.
* VNST, the data is neither normally distributed nor skewed.



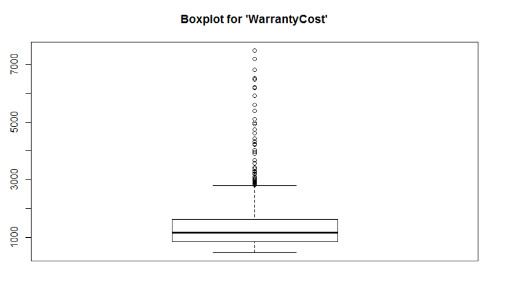
* MMR prices are all discrete and continuous values which will convey market value.
* ISonlinesale may have an impact with the perception that online purchases are more convenient for users.
* Vehodo follows the Gaussian distribution



* Vehage follows normal distribution
* Vechcost was right skewed and outliers can be sampled to get normally distributed values.



* Warranty cost is right skewed.



### Describe your methodology: give flowcharts, diagrams, pseudocode or formulas where appropriate.

For all the algorithms and methods implemented, we have done 4 types of sampling

1. Sampling with replacement: When a dataset row can be selected more than one time, we are doing **sampling with replacement.**
2. Sampling without replacement: When a dataset row can be selected only one time, we are doing **sampling without replacement**.
3. Stratified sampling with replacement: It divides the given dataset into smaller groups known as strata. All the strata are mutually exclusive. Then a sample is selected from each strata and same sample can be selected more than one time from each stratum.
4. Stratified sampling without replacement: when a sample from strata can be selected only one time, this is know as stratified sampling without replacement.

Stratified sampling provide a more accurate representation of the population based on the characteristics used to divide the population into strata.

### Evaluation Strategy:

For the evaluation of the methods implemented, we modified the algorithms not to only return the class of the data point but also the confidence value for each class. As we have confidence values for the each class. As a result, now we can calculate various measures easily.

We calculated the following measures:

* 1. Cross-validation: It is used to check the accuracy as kaggle did not provide test data with true labels, thus we used cross validation techniques to calculate the accuracy of the results provided by the classifier we used.
  2. Simple accuracy: Simple accuracy is summation of count of true positive and true negatives; it will give us crude measure of the accuracy.
  3. Balanced Accuracy: Balanced Accuracy is arithmetic mean of sensitivity and specificity.
  4. F1 Measures: Is the measure of test’s accuracy and it is a weighted harmonic mean of Recall and Precision.
  5. Precision – Recall: Precision is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved.
  6. ROC Curve: Is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

Baseline:

Right now, for this specific problem, the majority class count will be our baseline. The majority class is Not Bad Buy (i.e., value 0). This will be our simple accuracy based baseline.

Base Line = 87.7% (Not Bad Buy)

# Results:

## Preprocessing:

## Classification:

### Decision Tree:

## Sampling:

We implemented four methods for sampling:

1. Sampling without replacement
2. Sampling with replacement
3. Stratified sampling without replacement
4. Stratified sampling with replacement

We performed sampling over 10 different sample sizes, using above four techniques. Sample sizes were taken 10%, 20%, ……, 100% of the size of original data set.

We found the sampling with replacement tends to perform better than sampling without replacement, Also, found that simple sampling tend to perform better than stratified sampling.

Following are the figures for K-means classifier, for 10 different sample size with each of four sampling methods.

|  |  |
| --- | --- |
|  |  |
| C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_36490_3.png | C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_29190_3.png |
| C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_14590_3.png | C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_43780_3.png |
| C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_65680_3.png | C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_72980_3.png |
| C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_51080_3.png | C:\Users\user.local\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Accuracy_58380_3.png |

# Conclusion:

# Individual Tasks:

## Sachin Sable

## Mayank Jaglan

## Raghav Reddy

# References: