# TITLE: Am I Getting a Lemon?

## Problem Statement

One of the most important components of an auto dealership's business model is the ability to present customers with reliable vehicles that satisfy their needs. To do this companies in the auto selling industry must ensure that there are as few lemons as possible in their inventory (ideally they would have none). A lemon, or kick as it is also commonly referred to, is a vehicle purchased at an auction that has serious issues which prevent it from being re-sold. These problems often go unforeseen and are only realized after purchase. Once they are identified the car must be fixed or scrapped which can result in steep costs for the dealership (i.e., costs for transportation, throw-away repair work, market losses in reselling, etc.).

The purpose of this project is to develop a model that will predict which vehicles up for sale at an auction are at a high risk of being kicks. By doing this the model will lower the susceptibility of auto businesses to costs incurred from kicked cars, and thereby increase their ability to efficiently identify and assemble a quality inventory.

## Background and Related Work

This is essentially a class imbalance binary classification problem and many approaches can be applied. In this study Naïve Bayes' classifier, SVM, AdaBoost, Apriori Principle and Decision Trees are implemented and results are compared to evaluate which classifier performs better.

## Dataset

"Caravana" auto dealership dataset will be used for creating prediction model. The data contains 32 independent variables. These attributes are defined in the Data Dictionary in Appendix A.

The data set also includes binary class label i.e."IsBadBuy" that defines whether or not the vehicle was actually a kick. Additionally, some pieces of data are missing. The missing information will be handled appropriately in the data preparation step.

# Approach

## Data-Preprocessing and Exploration

Carvana dataset has 72983 records and out of that 8976 has positive samples and remaining were negative samples that makes it class imbalance problem. There are less positive samples available. Another issue with this dataset is that there was also missing data for a number of attributes.

### Data Normalization:

Carvana dataset has both numerical (such as prices), categorical (such as Make, Model etc) attributes. Categorical attributes are converted into numerical values.

Prices in dataset are normalized using min-max normalization technique between 1 and 100.

### Feature Selection:

Carvana dataset has 32 features and some of them have a lot of missing values. As a first step of data preprocessing features are reduced from 32 to 8 features:

1. Out of those 32 features based on analysis few attributes were removed such as record Id(RefId), Buyer Id etc as they do not contribute to classification. Following features were removed as a first step of feature selection: REfId,  PurchaseDate, Auction, VehYear, Nationality , BYRNO ,VNZIP ,VNST ,WheelTypeId.
2. Features PRIMEUNIT, AUCGUART has majority values as NULL so these are also removed for classification.
3. To identify the correlation between attributes a correlation matrix is built. Correlation values were not high so used cutoff value = 0.2. Following 13 feature were chosen based on correlation values:
   VehicleAge, Make, SubModel, VehOdo, MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitonRetailCleanPrice,MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice, VehBCost

4. Then used backward elimination(AIC) to eliminate more features and below are final 8 features used for classification:
    VehicleAge ,Make,
    SubModel ,VehOdo,
    MMRAcquisitionAuctionAveragePrice,
    MMRAcquisitionAuctionCleanPrice,
    MMRAcquisitonRetailCleanPrice ,VehBCost

## Missing Values:

In caravan dataset some records have missing values for a lot of features and they were mostly categorical attributes. For the sake of simplicity records with missing values are removed that reduced the dataset to 69804 total records. 63065 negative samples and 6739 positive samples.

## Model Building and Evaluation

To predict whether an auction is a kick or not various classification models are built using Naïve Baye's classifier, AdaBoost and SVM and best rules are also found using Apriori principle.

In all models same 8 features were used with 5 fold cross validation.

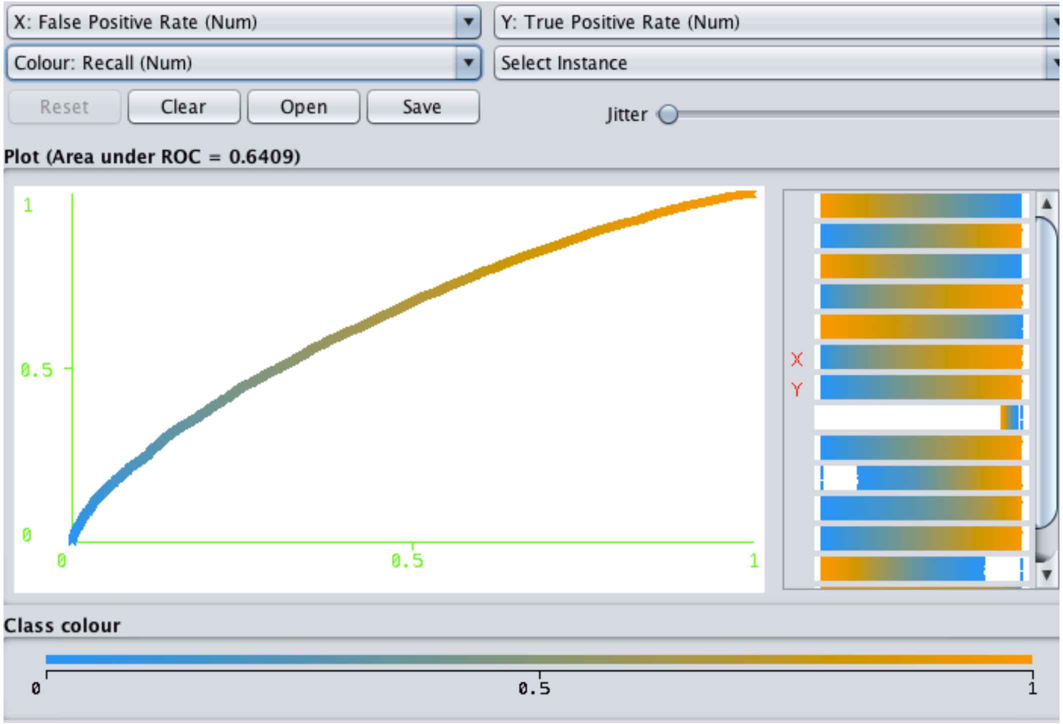
## Naïve Baye's Classifier Results

Confusion matrix:

|  |  | Predicted |  |
|---|---|---|---|
|  |  | 0 | 1 |
| Actual | 0 | 55410 | 7655 |
|  | 1 | 5068 | 1671 |

Model Accuracy = 81.77%

ROC Curves are below and area under the curve is 0.6409



X: False Positive Rate (Num)

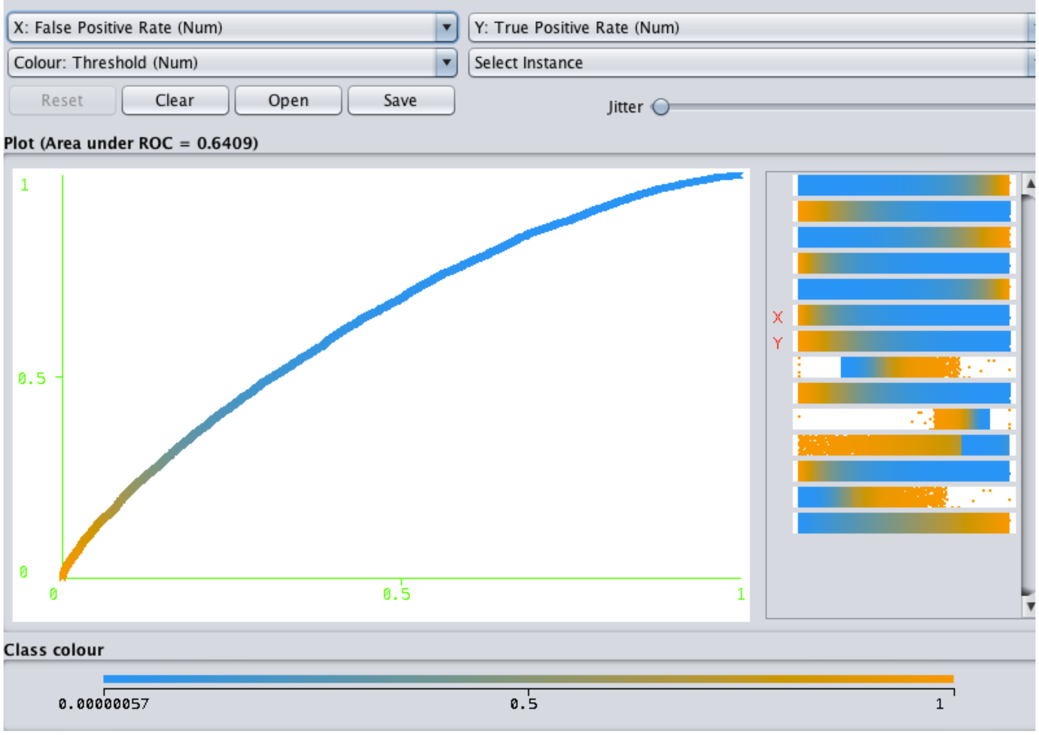
Y: True Positive Rate (Num)

Colour: Recall (Num)

Select Instance

Plot (Area under ROC = 0.6409)

Figure 1- ROC for class-0 (Naive Baye's)



X: False Positive Rate (Num)

Y: True Positive Rate (Num)

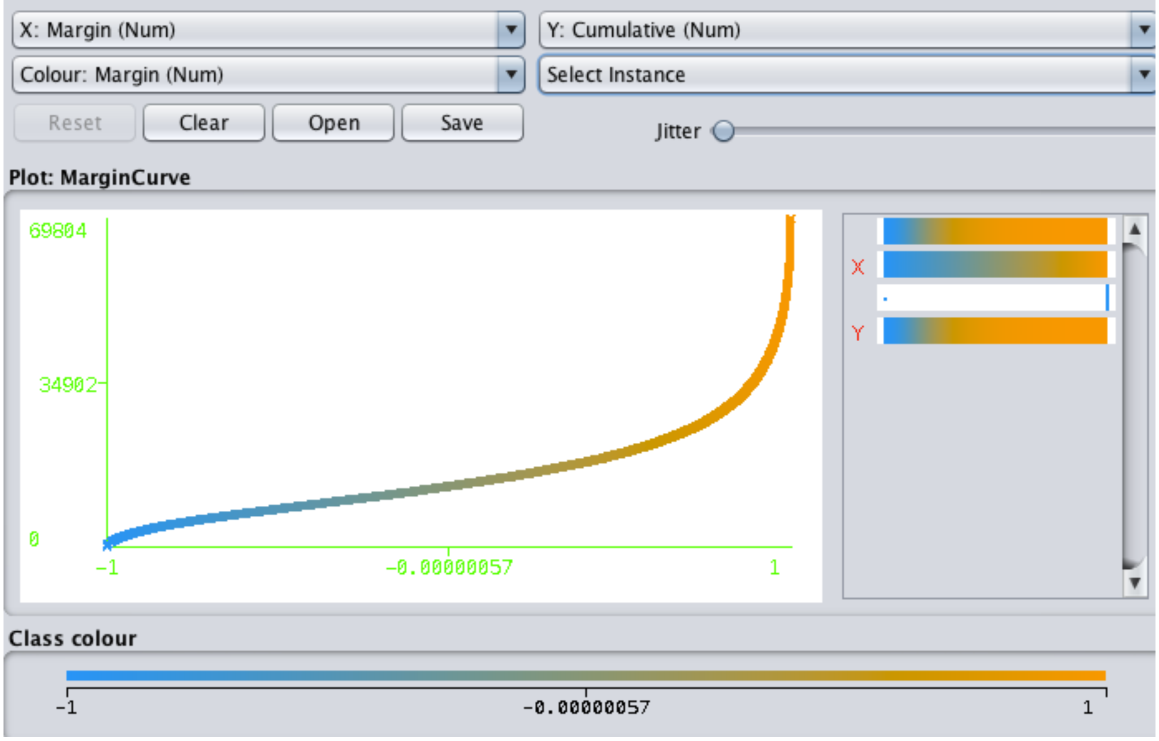
Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.6409)

Figure 2- ROC for class-1 (Naive Baye's)

Below is the margin curve for Naïve Bayes classifier



X: Margin (Num)

Y: Cumulative (Num)

Colour: Margin (Num)

Select Instance

Figure 3- Margin Curve for Naive Baye's

## AdaBoost Results

Adaboost algorithm used VehicleAge as Decision Stump.

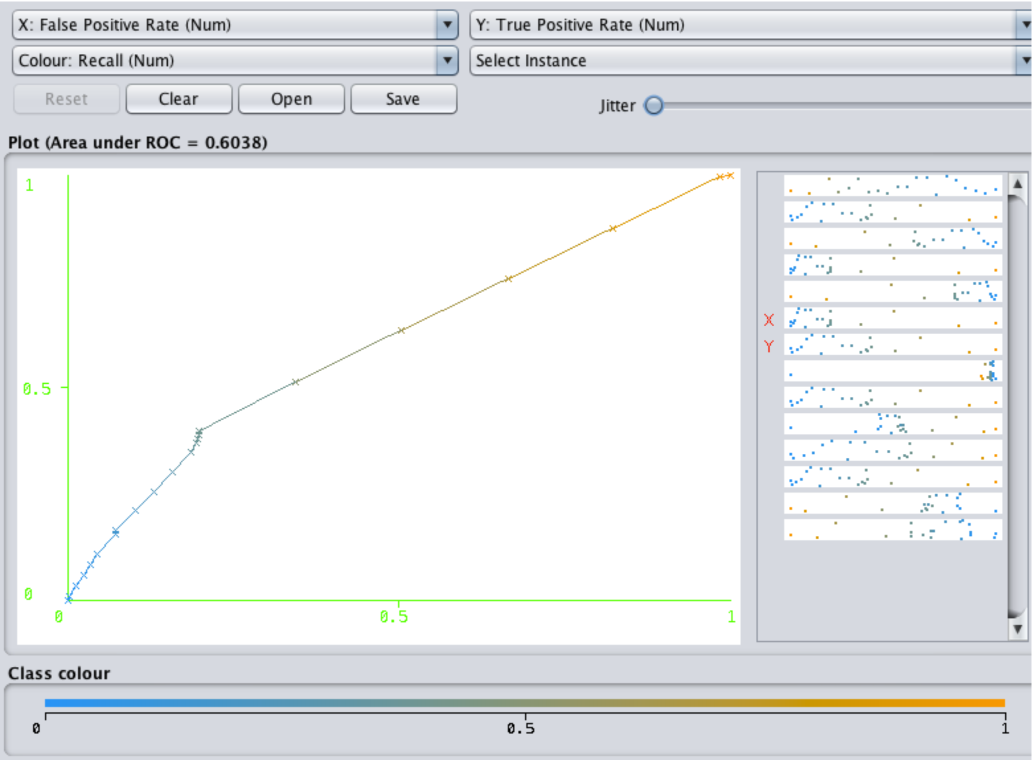
Confusion matrix:

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | 63065 | 0 |
| | 1 | 6739 | 0 |

Model Accuracy = 90.3458%

Though model accuracy is higher but it did not classify any of the positive samples correctly.

ROC Curves are below and area under the curve is 0.60



X: False Positive Rate (Num)

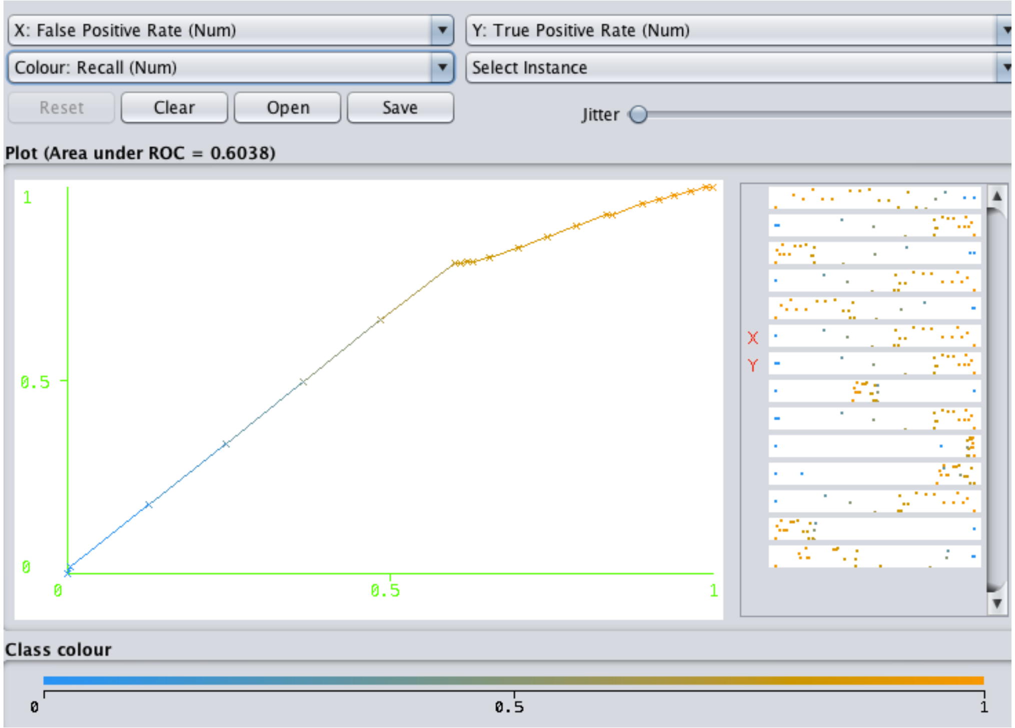
Y: True Positive Rate (Num)

Colour: Recall (Num)

Select Instance

Plot (Area under ROC = 0.6038)

Figure 4- ROC curve for class-0 ( Adaboost)



X: False Positive Rate (Num)

Y: True Positive Rate (Num)

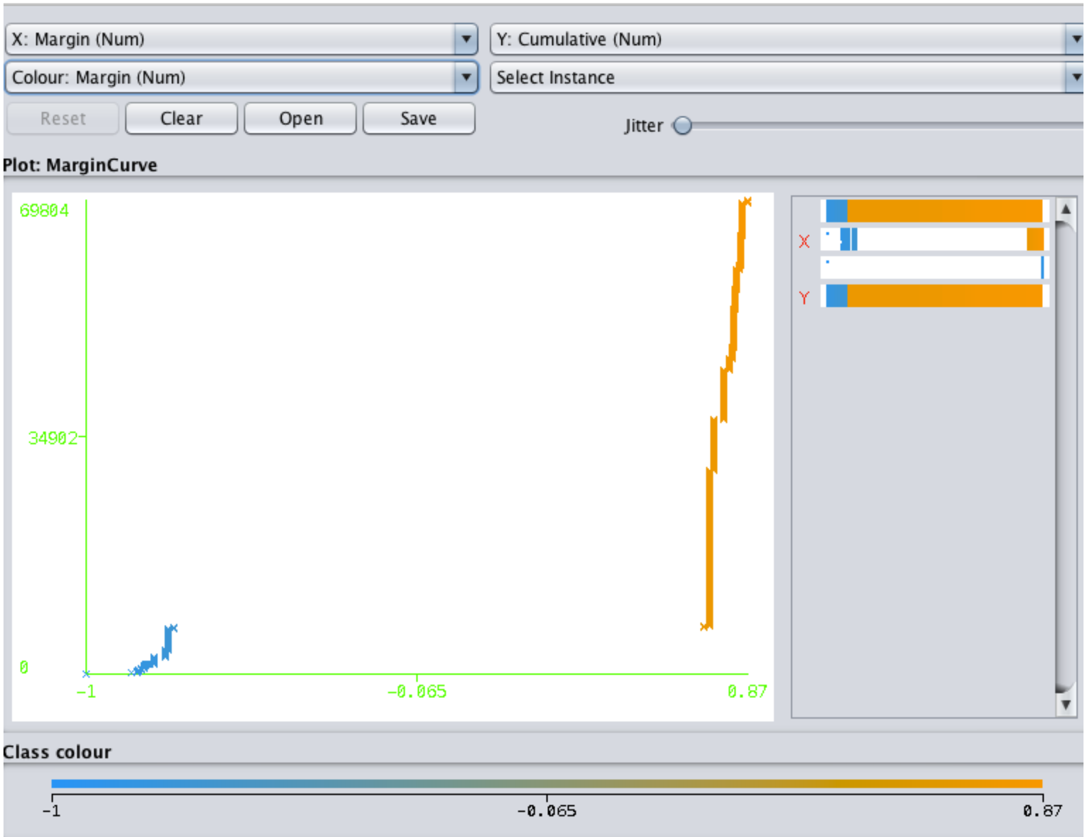
Colour: Recall (Num)

Select Instance

Plot (Area under ROC = 0.6038)

Figure 5- ROC for class-1 (Adaboost)

Below is the margin curve for Adaboost



X: Margin (Num)

Y: Cumulative (Num)

Colour: Margin (Num)

Plot: MarginCurve

Figure 6- Margin Curve for Adaboost

## SVM

Confusion matrix for SVM is given by

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | 63065 | 0 |
| | 1 | 6739 | 0 |

Model Accuracy = 90.3458%

Though accuracy for SVM is higher but it did not classify any of the positive samples correctly.

## Apriori Results

Used Apriori to find out interesting association rules. Below are 5 most interesting rules based on confidence and lift:

| Rules | Confidence | Lift |
|---|---|---|
| VehicleAge=2 ==> IsBadBuy=0 | 0.96 | 1.06 |
| VehicleAge=3 ==> IsBadBuy=0 | 0.94 | 1.04 |
| Make=4 ==> IsBadBuy=0 | 0.92 | 1.02 |
| Make=6 ==> IsBadBuy=0 | 0.92 | 1.02 |
| VehicleAge=4 ==> IsBadBuy=0 | 0.92 | 1.02 |

None of the rules were generated for class=1.

## Decision Trees

'VehicleAge' feature was also used as root node in decision stump.

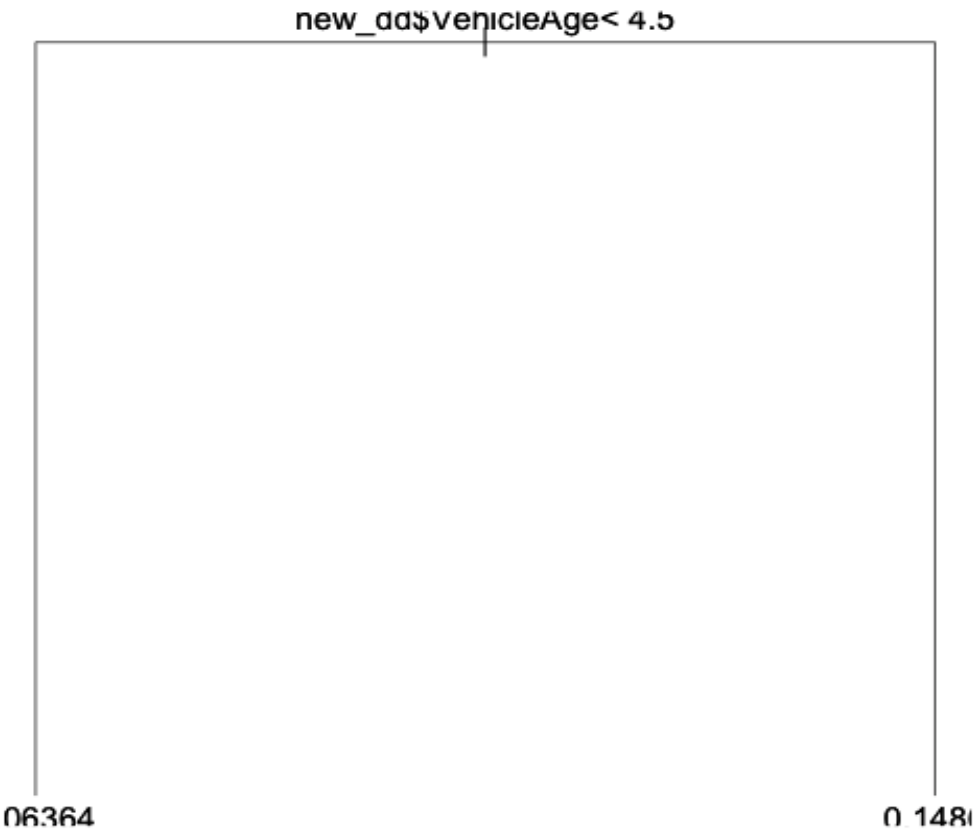


Figure 7- Decision Tree with one stump (VehicleAge)

## Code:

I have used R and weka tool for data preprocessing and models. Files can be found at below link: https://www.dropbox.com/sh/qggtvlz64iagrzf/AAC9Wua0HoVaDQmybajSCWrZa?dl=0

## Conclusion:

Carvana has imbalance class data having 90% negative samples and 10% positive samples. Among Naïve Baye's, SVM, AdaBoost classifiers the Naïve Baye's has good results. Though accuracy is higher for SVM and Adaboost but they do not cover positive samples at all whereas Naïve Bayes' has covered both samples.
Apriori did not find any rule for class label=1 due to imbalance nature of data. From Apriori and Decision Trees results its concluded that among all features 'VehicleAge' is important attribute for classification.

It can be concluded that Naïve Baye's has better performance for imbalance class problem and 'VehicleAge' is important attribute for kicked vehicle classification.

## References

- Carvana: https://www.carvana.com/
- Weka: http://www.cs.waikato.ac.nz/ml/weka/
- R: https://www.r-project.org

## Appendix A: Data Dictionary

Taken from the Carvana_Data_Dictionary

| FIELD NAME | DEFINITION |
|---|---|
| RefID | Unique (sequential) number assigned to vehicles |
| IsBadBuy | Identifies if the kicked vehicle was an avoidable purchase |
| PurchDate | The date the vehicle was purchased at auction |
| Auction | Auction provider at which the vehicle was purchased |
| VehYear | The manufacturer's year of the vehicle |
| VehicleAge | The years elapsed since the manufacturer's year |
| Make | Vehicle manufacturer |
| Model | Vehicle model |
| Trim | Vehicle trim level |
| SubModel | Vehicle sub-model |
| Color | Vehicle color |

| | |
|---|---|
| Transmission | Vehicles transmission type (automatic, manual) |
| WheelTypeID | The type ID of the vehicle wheel |
| WheelType | The vehicle wheel type description (alloy, covers) |
| VehOdo | The vehicles odometer reading |
| Nationality | The manufacturer's country |
| Size | The size category of the vehicle (compact, SUV, etc.) |
| TopThreeAmericanName | Identifies if the manufacturer is one of the top three American manufacturers |
| MMRAcquisitionAuctionAveragePrice | Acquisition price for this vehicle in average condition at time of purchase |
| MMRAcquisitionAuctionCleanPrice | Acquisition price for this vehicle in the above Average condition at time of purchase |
| MMRAcquisitionRetailAveragePrice | Acquisition price for this vehicle in the retail market in average condition at time of purchase |
| MMRAcquisitonRetailCleanPrice | Acquisition price for this vehicle in the retail market in above average condition at time of purchase |
| MMRCurrentAuctionAveragePrice | Acquisition price for this vehicle in average condition as of current day |
| MMRCurrentAuctionCleanPrice | Acquisition price for this vehicle in the above condition as of current day |
| MMRCurrentRetailAveragePrice | Acquisition price for this vehicle in the retail market in average condition as of current day |
| MMRCurrentRetailCleanPrice | Acquisition price for this vehicle in the retail market in above average condition as of current day |
| PRIMEUNIT | Identifies if the vehicle would have a higher demand than a standard purchase |
| AUCGUART | The level guarantee provided by auction for the vehicle (green light - guaranteed/arbitratable, yellow light - caution/issue, red light - sold as is) |
| BYRNO | Unique number assigned to the buyer that purchased the vehicle |
| VNZIP | Zip code where the car was purchased |
| VNST | State where the car was purchased |
| VehBCost | Acquisition cost paid for the vehicle at time of purchase |
| IsOnlineSale | Identifies if the vehicle was originally purchased online |
| WarrantyCost | Warranty price (term = 36 month  and millage = 36K) |