# Don’t get kicked

Data College | Capstone Project of Machine Learning Case 1

**Mayuko Tahara, F&C Japan**

## 1 Table of Contents

# 2 PROBLEM OVERVIEW

"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. The challenge of this competition is **to predict if the car purchased at the Auction is a Kick (bad buy).**" It means there are only two outcomes, hence this is a classification problem. Target indicator in data is a binary field "IsBadBuy" which indicates the car purchased at the auction is a bad buy or not.

# 3 DATA DESCRIPTION

## 3.1 DATA FIELDS

There are 34 fields in total. Among them, 25 fields are string and 9 fields are numbers.

| Category | Field Name | Data Type | Definition |
|---|---|---|---|
| Unique Identifier | RefID | Number (integer) | Unique (sequential) number assigned to vehicles |
| Target Field | IsBadBuy | Number (integer) | Identifies if the kicked vehicle was an avoidable purchase |
| Auction Info | PurchDate | String | The Date the vehicle was Purchased at Auction |
| | Auction | String | Auction provider at which the vehicle was purchased |
| Vehicle Info | VehYear | Number (integer) | The manufacturer's year of the vehicle |
| | VehicleAge | Number (integer) | The Years elapsed since the manufacturer's year |
| | Make | String | Vehicle Manufacturer |
| | Model | String | Vehicle Model |
| | Trim | String | Vehicle Trim Level |
| | SubModel | String | Vehicle Submodel |
| | Color | String | Vehicle Color |
| | Transmission | String | Vehicles transmission type (Automatic, Manual) |
| | WheelTypeID | String | The type id of the vehicle wheel |
| | WheelType | String | The vehicle wheel type description (Alloy, Covers) |
| | VehOdo | Number (integer) | The vehicles odometer reading |
| | Nationality | String | The Manufacturer's country |
| | Size | String | The size category of the vehicle (Compact, SUV, etc.) |
| | TopThreeAmericanName | String | Identifies if the manufacturer is one of the top three American manufacturers |

| | | | |
|---|---|---|---|
| Acq Price Info | MMRAcquisitionAuctionAveragePrice | String | Acquisition price for this vehicle in average condition at time of purchase |
| | MMRAcquisitionAuctionCleanPrice | String | Acquisition price for this vehicle in the above Average condition at time of purchase |
| | MMRAcquisitionRetailAveragePrice | String | Acquisition price for this vehicle in the retail market in average condition at time of purchase |
| | MMRAcquisitonRetailCleanPrice | String | Acquisition price for this vehicle in the retail market in above average condition at time of purchase |
| | MMRCurrentAuctionAveragePrice | String | Acquisition price for this vehicle in average condition as of current day |
| | MMRCurrentAuctionCleanPrice | String | Acquisition price for this vehicle in the above condition as of current day |
| | MMRCurrentRetailAveragePrice | String | Acquisition price for this vehicle in the retail market in average condition as of current day |
| | MMRCurrentRetailCleanPrice | String | Acquisition price for this vehicle in the retail market in above average condition as of current day |
| Risk Info | PRIMEUNIT | String | Identifies if the vehicle would have a higher demand than a standard purchase |
| | AUCGUART | String | The level guarantee provided by auction for the vehicle (Green light - Guaranteed/arbitratable, Yellow Light - caution/issue, red light - sold as is) |
| Unique Identifier | BYRNO | Number (integer) | Unique number assigned to the buyer that purchased the vehicle |
| Geo Info | VNZIP | String | Zipcode where the car was purchased |
| | VNST | String | State where the car was purchased |
| Others | VehBCost | Number (double) | Acquisition cost paid for the vehicle at time of purchase |
| | IsOnlineSale | Number (integer) | Identifies if the vehicle was originally purchased online |
| | WarrantyCost | Number (integer) | Warranty price (term = 36month and millage = 36K) |

## 3.2 Descriptive Statistics

Data is loaded using "Statistics" node in KNIME in order to grasp an overview about the data. Here are the key findings:

(1) There are 72983 entries in train data and 48000 entries in test data
(2) "RefId", "BYRNO" are unique identifiers for vehicles/buyers therefore not related to this problem. They can be removed.
(3) "PRIMEUNIT" and "AUCGUART" have too many null values (96%=69564/72983) therefore they can be dropped too.

| PRIMEUNIT | AUCGUART |
|---|---|
| No. missings: 0 | No. missings: 0 |
| Top 20: | Top 20: |
| NULL : 69564 | NULL : 69564 |
| NO : 3357 | GREEN : 3340 |
| YES : 62 | RED : 79 |

*Figure 1. Descripting Analytics of PRIMEUNIT and AUCGUART*

(4) Other fields include few null values, but not as many as "PRIMEUNIT" and "AUCGUART"
(5) "VehYear" and "VehAge" are highly correlated as they both indicate how old the vehicle is. Since "VehAge" directly expresses its age, it is ideal to keep this one and disregard "VehYear."

# 4 EXPLORATORY DATA ANALYSIS

In this section, several data visualizations are done to explore relationships among fields by using "Interactive Histogram", "Pivoting', "Scatter Plot", "Interactive Pie Chart" and "Linear Correlation."

## 4.1 UNIVARIATE ANALYSIS

Univariate histogram shows distinctive distribution for some fields.

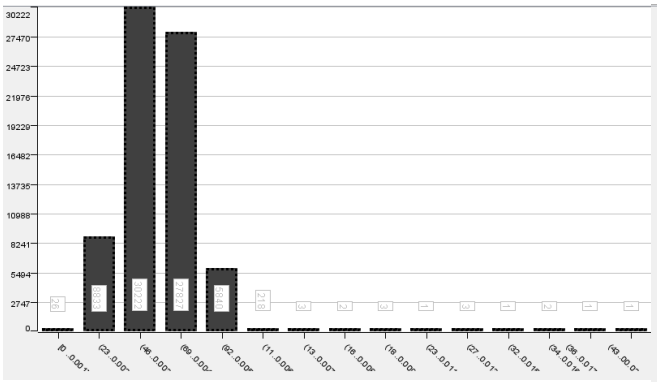

*Figure 2. Distribution of VehBCost*

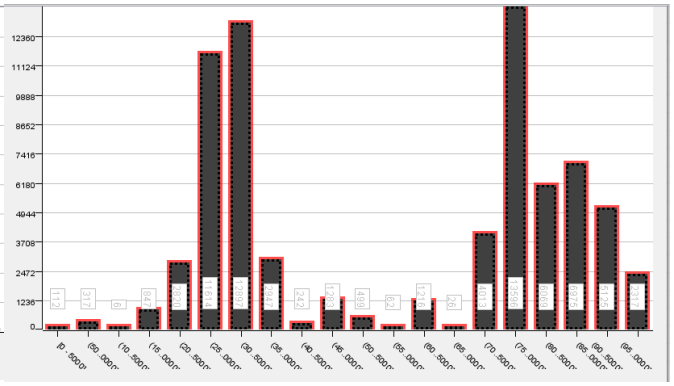

*Figure 3. Distribution of VNZIP*

VehB cost is an actual amount that is paid to acquire vehicles. Data is concentrated on 4600-6900 band and very few on 6900- band. It is interesting to see that acquisition price is extremely high for several cases. VNZIP is a zip code where cars are purchased and there are two peaks in distribution – which indicates that there are two big auction marketplaces.

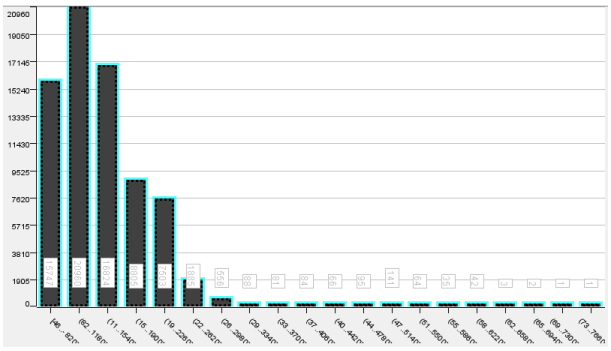

*Figure 4. Distribution of Warranty Cost*

Warranty cost distribution is left skewed which presents that insignificant amount of warranty is paid in most cases.

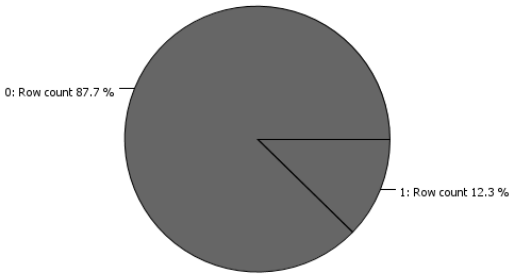


*Figure 5. Pie Chart of IsBadBuy*

Figure 5 presents a distribution of IsBadBuy, a target predictor. It shows 12.3% of total cars are kicked.

## 4.2 Correlation Matrix

The correlation matrix highlights several key findings.

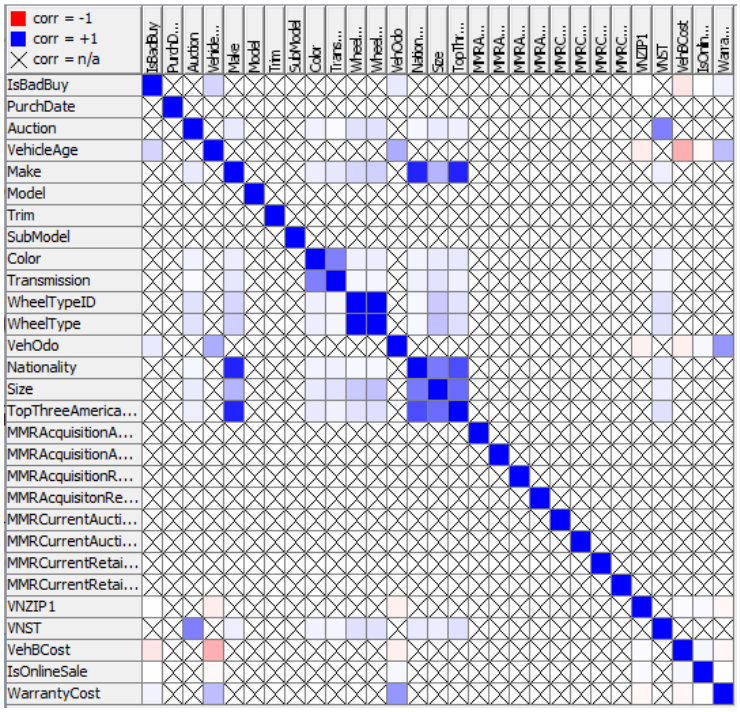


*Figure 6. Correlation Matrix*

(1) There are very strong correlation between "Make" and "Nationality" and "Make" and "TopThreeAmericans." Here we will keep "Nationality" and drop others.
(2) "WheelTypeID" and "WheelType" are highly correlated as they both indicate the same information. Therefore "WheelTypeID" is better to be taken out.
(3) "Nationality" has strong correlation with "Size"
(4) "VNST" has strong correlation with "Auction."
(5) "WarrantyCost" has strong correlation with "VehOdo". This means that warranty gets more expensive as distance travelled gets longer.
(6) The target variable "IsBadBuy" has strong correlation with "VehicleAge", "VehOdo", "VehBCost" and "WarrantyCost." It is worthwhile to look at them in details.

## 4.3 Bivariate Analysis

Deep dive into fields that showed correlation with each other in 4.2 gave us hints on reasons why they are related.

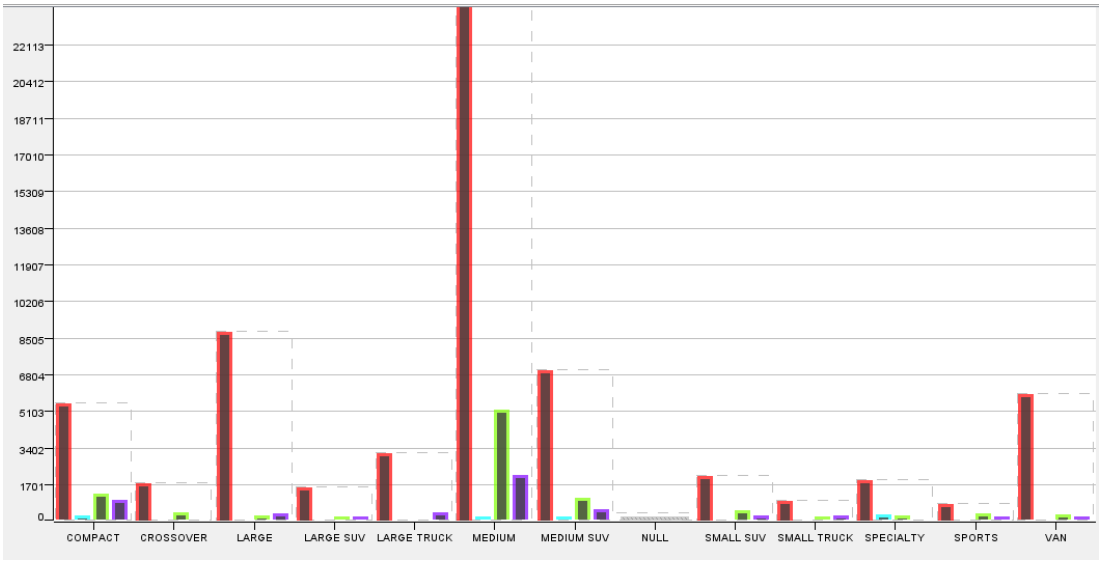


*Figure 7. Number of Cars by Nationality and Size*
*(Red=American, Green=Other Asian, Purple=Top Line Asian, Blue=Other)*

In general, American cars is majority in all car size. It is especially dominant in LARGE, LARGE TRUCK and VAN. Car size of Non-American brands at auction are mostly MEDIUM, MEDIUM SUV or COMPACT.

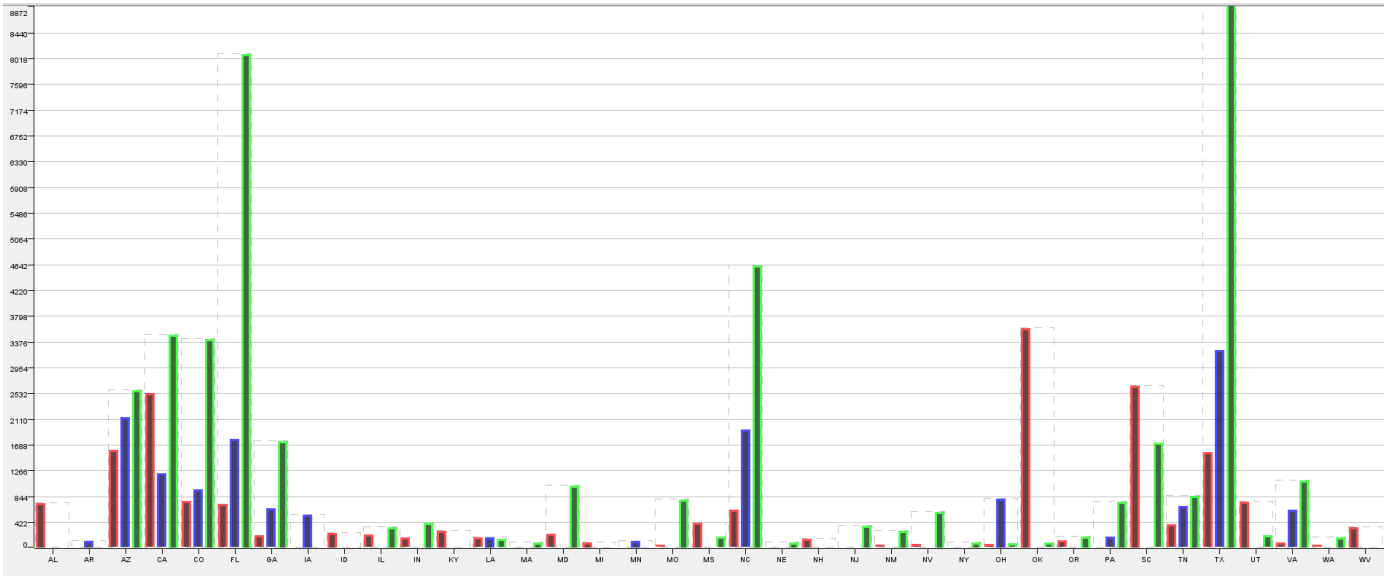


*Figure 8. Number of Cars By VNST and Auction (Green=MANHEIM, Blue=ADESA, Red=Other)*

Figure 8 shows that MANHEIM, the world's largest auto auction company, is dominant in all the states except AL, OK and SC where "others" are dominant.

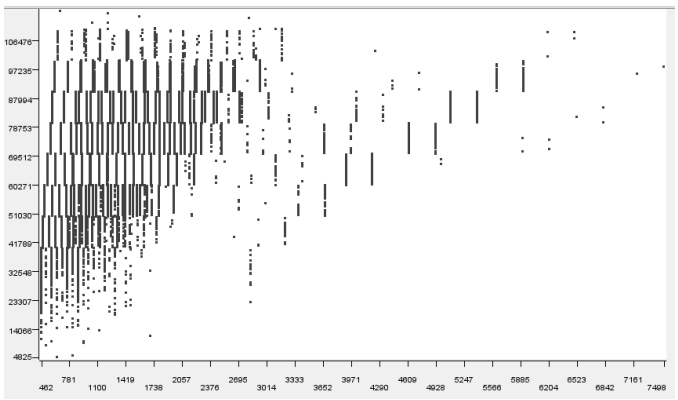


*Figure 9. Scatter Plot of Warranty Cost and VehOdo (X-axis: Warranty Cost, Y-axis: VehOdo)*

Figure 9 presents relationship between warranty costs and numbers in odometers (= total distance travelled). It shows relatively strong positive correlation. However, there are also cars that travelled distances are long but warranty is low.

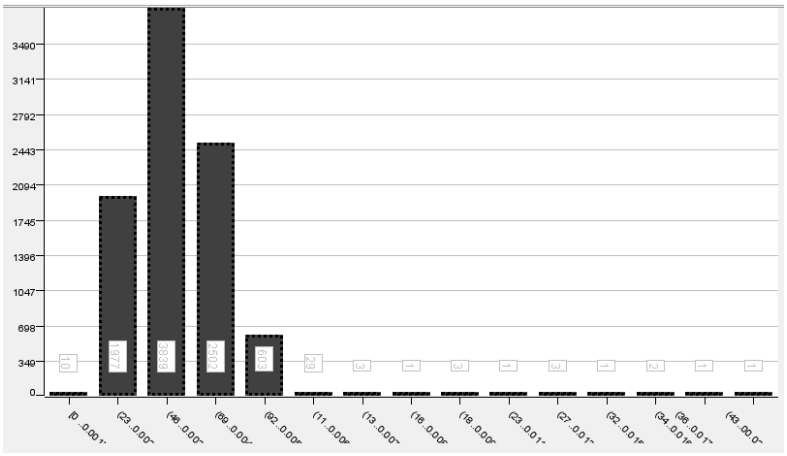


*Figure 10. Kicks by VehBcost*

Lastly, analysis between IsBadBuy and several features unveils further insights. Figure 10 shows a number of kicks by VehBCost. Comparing to Figure 2 that presents number of cars by VehBCost, number of total cars and kicked cars are almost the same after 11500 band. In other words, it is very likely to be a kick if acquisition cost is higher than 11500. What kind of cars are they?

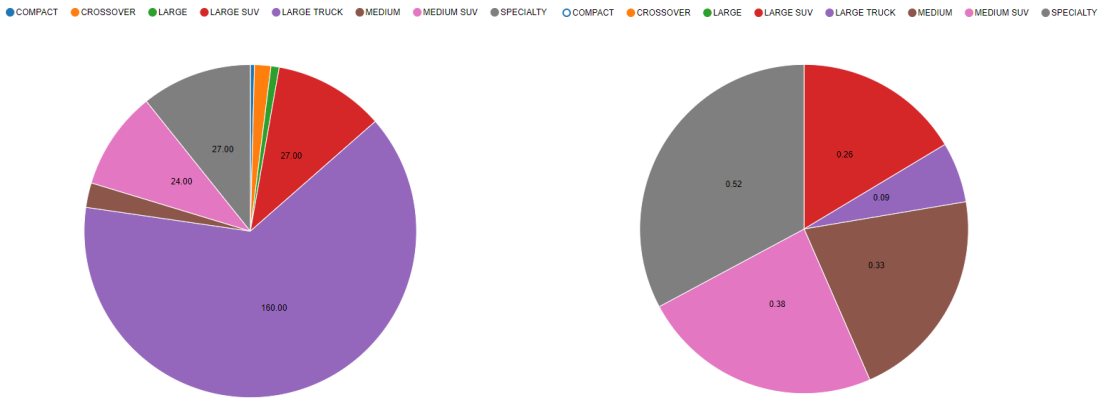


*Figure 11. Ratio of Number of Cars by Car Size* *Figure 12. Ratio of Probability of Kicked Cars by Car Size Where acquisition cost is equal to or more than 11500*

Looking at filtered data with condition "VehBCost >= 11500", it looks different than in Figure 7 which shows number of cars by car size using all data. Unlike Figure 7, large truck is a majority with the condition. However, specialty is most likely kicked among these cars.

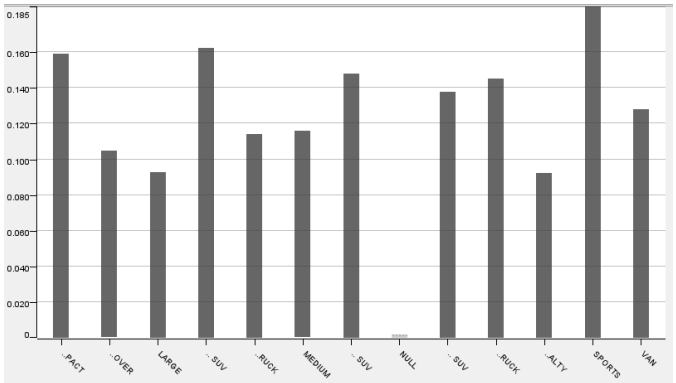


*Figure 13. Probability of Kicked Cars by Car Size*

This is contrary to a finding presented in Figure 13 which shows a probability of being kicked by car size using all data. Here, sports cars are most likely to be kicked and specialty cars are less likely. In a nutshell, sports cars in general are prone to be bad due to mechanical stress, but specialty is the one that can be a lemon in some cases where acquisition cost is high.

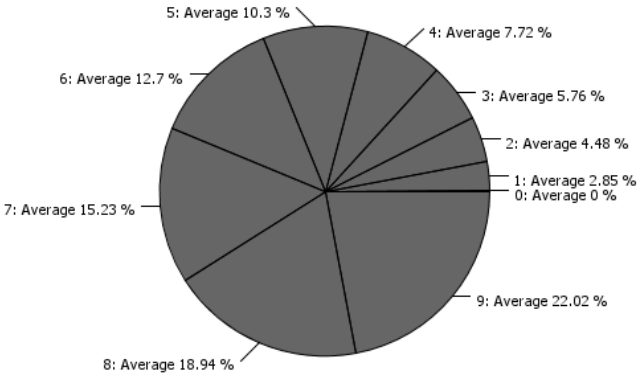


*Figure 14. Probability of Being Kicked By Vehicle Age*

With vehicle age, the finding is straightforward. The probability of being kicked cars is getting higher as vehicles get older.

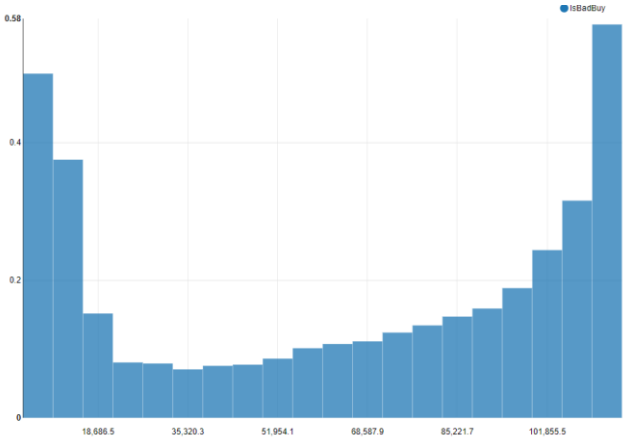


*Figure 15. Probability of Being Kicked By VehOdo (X-axis: VehOdo, Y-axis: probability of being kicked)*

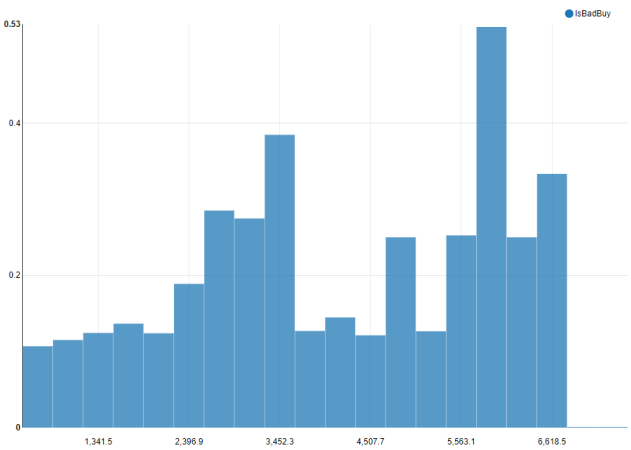


*Figure 16. Probability of Being Kicked By Warranty Cost (X-axis: Warranty cost, Y-axis: Probability of being kicked)*

Odometer can be treated as an indirect indicator of vehicle age, but Figure 15 presents different distribution from Figure 14. The probability of being kicked for cars with extremely low odometers is almost as the same as the probability of being kicked for cars with very high odometers.

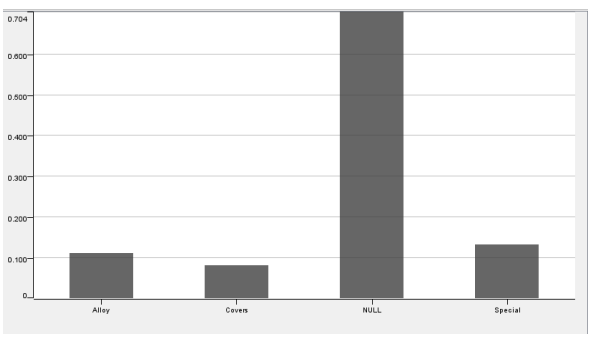


*Figure 17. Probability of Being Kicked By WheelType*

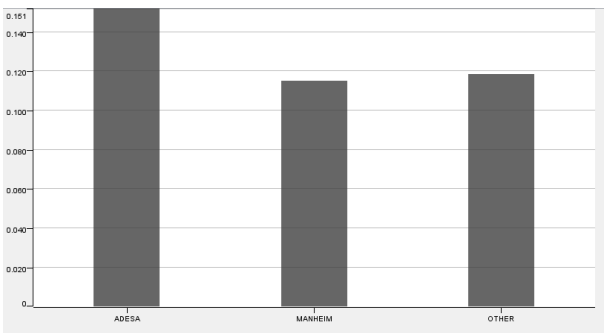


*Figure 18. Probability of Being Kicked By Auction*

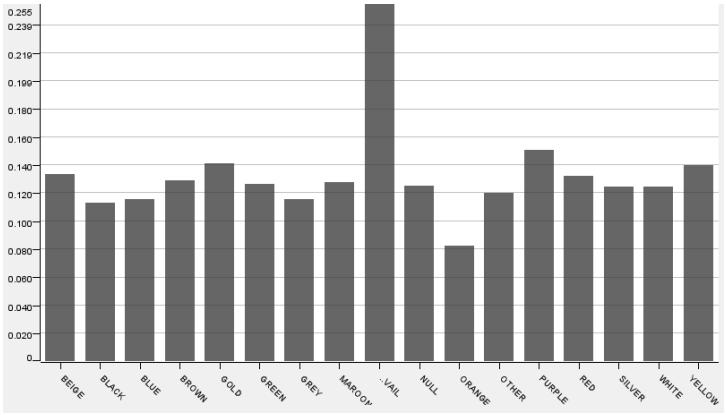


*Figure 19. Probability of Being Kicked By Color*

Is there any other factors that are related to IsBadBuy which are not shown in correlation matrix? After a quick check, several analysis return interesting results. Null value in wheel type and color is much more likely to be a lemon than other categories, therefore it should be treated as a new category. In terms of Auction, "ADESA" is most likely to be kicked. Recalling figure 8, "ADESA" is not a common auto auction company except in some states.

# 5 Feature Engineering

In this section, we will select variables and explain how to transform them for prediction based on section 3 and 4.

| Field Name | Data Type | Transform note |
|---|---|---|
| IsBadBuy | Integer | Convert to string |
| Auction | String | |
| VehicleAge | Integer | |
| Nationality | String | |
| WheelType | String | Treat null as different category |
| VehOdo | Integer | |
| VNST | String | |
| VehBCost | Number (double) | |
| WarrantyCost | Integer | |
| Sports_Flag | Integer | 1 if SIZE is "Sports", else 0 |
| Color_Flag | Integer | 1 if COLOR = "NOT AVAIL" or null, else 0 |

Sport_flag, speciality_flag and color_flag are new features created based on findings of data exploration in previous section. Below figure shows how features are created or converted:

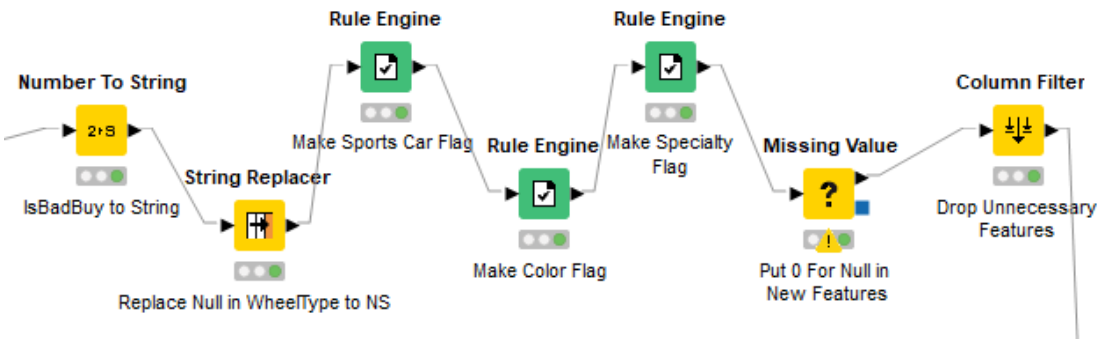


*Figure 20. Data Transformation*

Automatic feature selection is also done using Random Forest algorithm. As a result, highlighted features are selected:

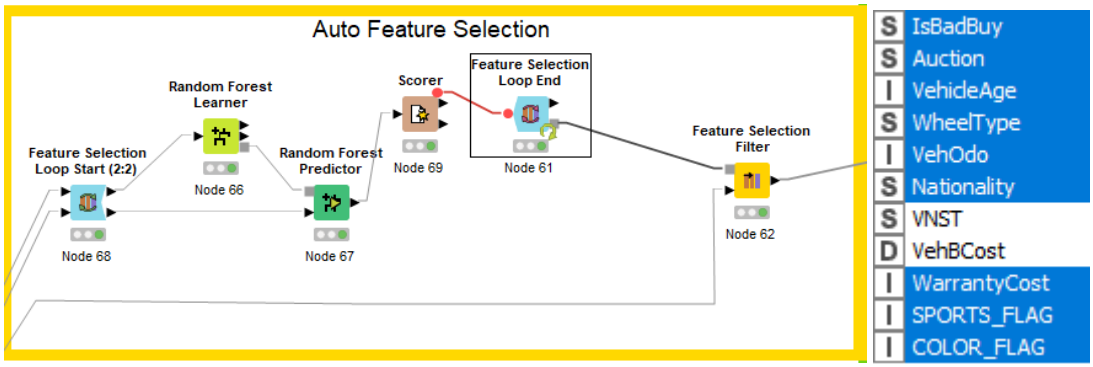


*Figure 21. Auto Feature Selection*

# 6 Data Modeling and Evaluation

In this section, we will sample training data and create 2 models to predict IsBadBuy and evaluate the result.

## 6.1 Create Balanced Data

As we found out in figure 5, there are far more good buys data than bad buys which will result in low F1 and recall. Therefore we extract 15% of total good buys in order to balance good and bad buy data.

| S IsBadBuy | I Count (IsBadBuy) |
|---|---|
| 0 | 9601 |
| 1 | 8976 |

*Figure 22. Number of Data for Good and Bad Buy*

## 6.2 Model 1

In model 1, we will simply consider all the features except new features created (Sports_flag and Color_flag). Random forest is used as a classification algorithm and Gini index is chosen for a split criterion.

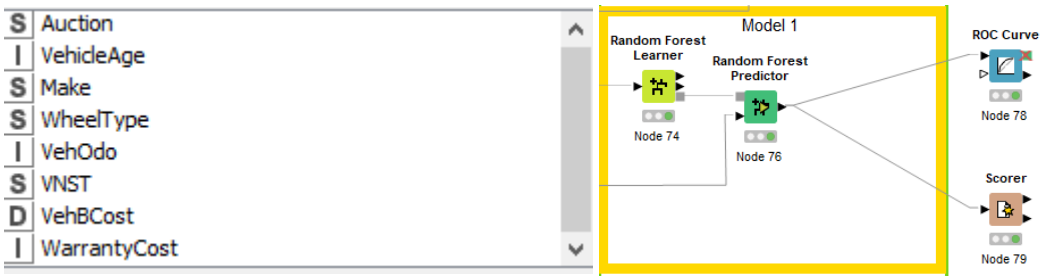


*Figure 23. List of Features Used and Modeling Steps (Model 1)*

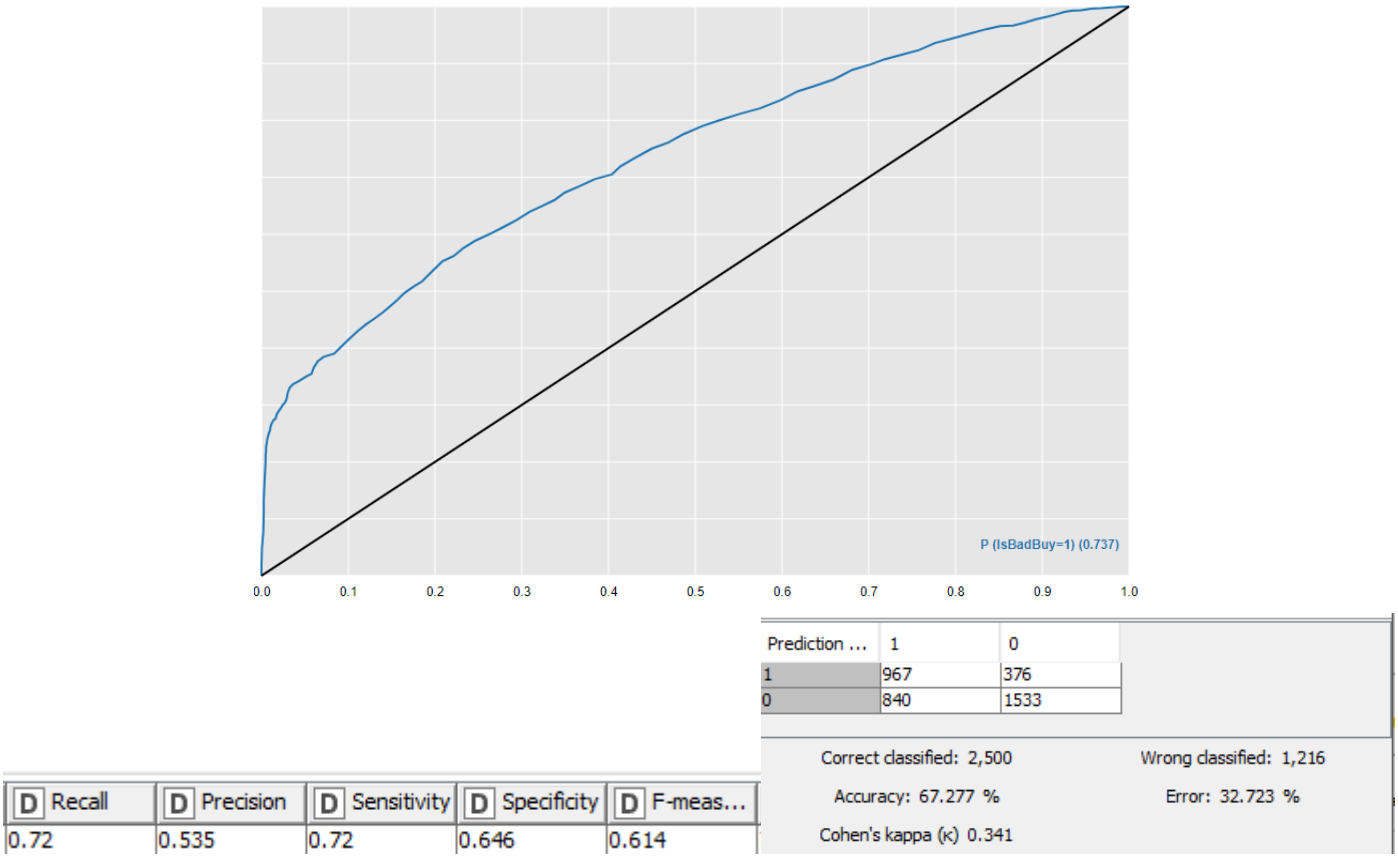


*Figure 24. Scores for Model 1*

## 6.3 Model 2

Model 2 is constructed based on auto feature selection results; highlighted features are used. Algorithm and scoring method are the same as model 1.

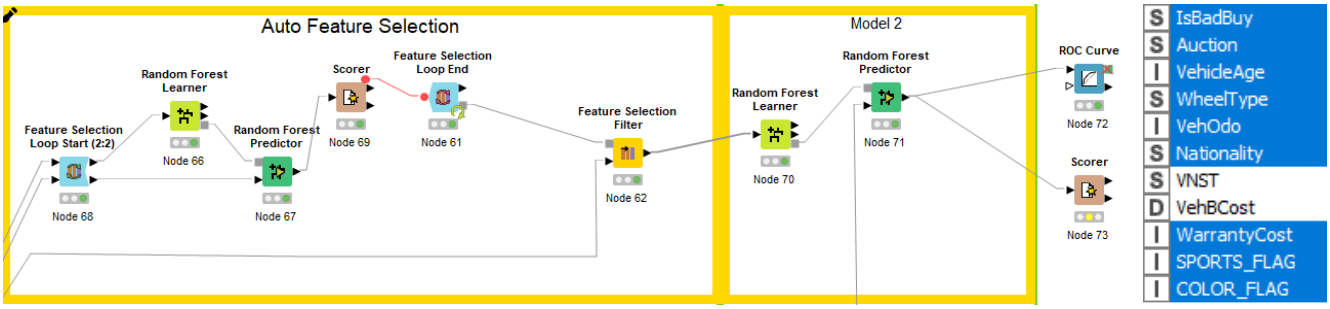


*Figure 25. List of Features Used and Modeling Steps (Model 2)*

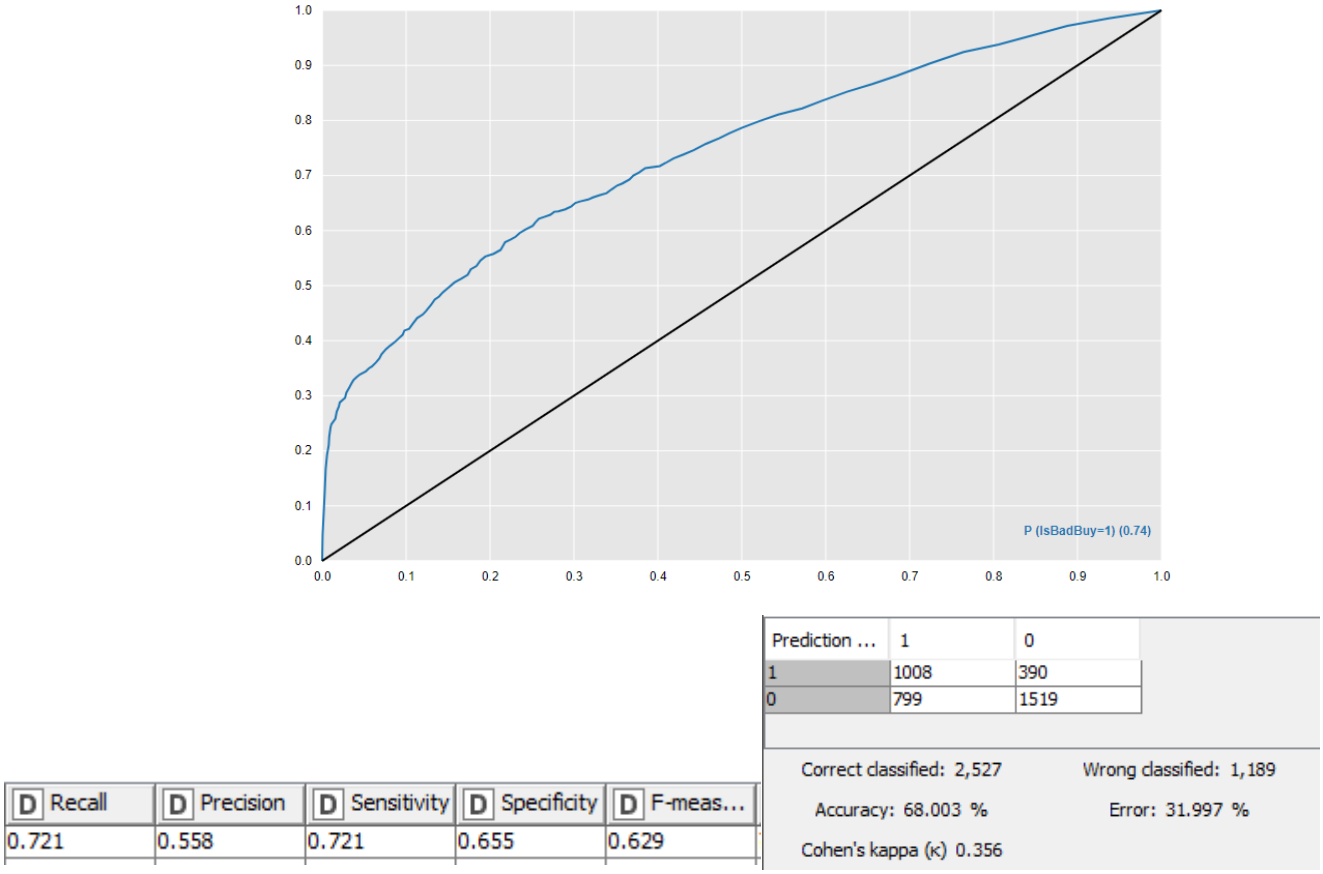


*Figure 26. Scores for Model 2*

Comparing 2 models, model 2 is slightly better in all scores; AUC, accuracy, recall, precision and F1.

## 6.4 Result and Summary

For a prediction submission to Kaggle, classifier trained in Model 2 is used on test dataset as it performs better than Model 1. As a result, it returned a public score of 0.14944 and ranked 394th.


*Figure 27. Scores in Kaggle*

In model 2, we considered two new features based on data exploration. Auto feature selection chooses them and this model returned better result than model 1 which doesn't use them. This outcome highlights the importance of exploratory data analysis with deep understanding of data and problem background.