Data Intensive Computing

Term Project – Tableau

Shruti Kulkarni	50207124
Kirti Hari	50208065

Introduction

In this project, an analysis using a chosen prediction learning model has been done to a dataset related to vehicle purchase made at auctions, this dataset was obtained from Kaggle competitions entitled "Do not Get Kicked!". The purpose of the project is to create a model to predict the vehicle

purchase be in the best condition possible, hence reduce the risk of buying a car in poor condition. We used several Machine Learning models for analysis but one in particular gave a big leap in the level of accuracy.

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.

Data Interpretation

As mentioned, we obtained the dataset from Kaggle.com challenge "Don’t Get Kicked" hosted by Carvana. The data set contains 32 unique features with 73,041 samples along with a labeling of 0 for good car purchases and 1 for "kicks" with the corresponding variable termed as “Is Bad Buy”. Some of the key features in the dataset are Vehicle cost, Odometer Readings, Age of the Vehicle, Make of the vehicle, etc.

On analyzing the data more deeply, we realized that the given dataset is skewed, with a staggering bias towards “good cars” which were over 80% of the data sample. Because of this, we could achieve only a certain level of precision and accuracy as opposed to an unbiased data sample.

For using the data to make our predictions, we needed it to be as clean as precise as possible. Hence we cleaned the data in R to make it as accessible as possible for further use.

Data Analysis

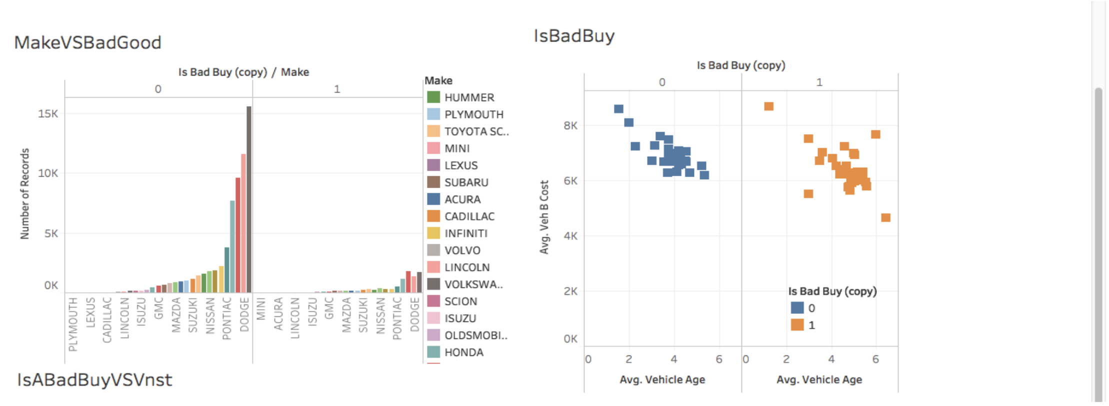
Analysis Scope:

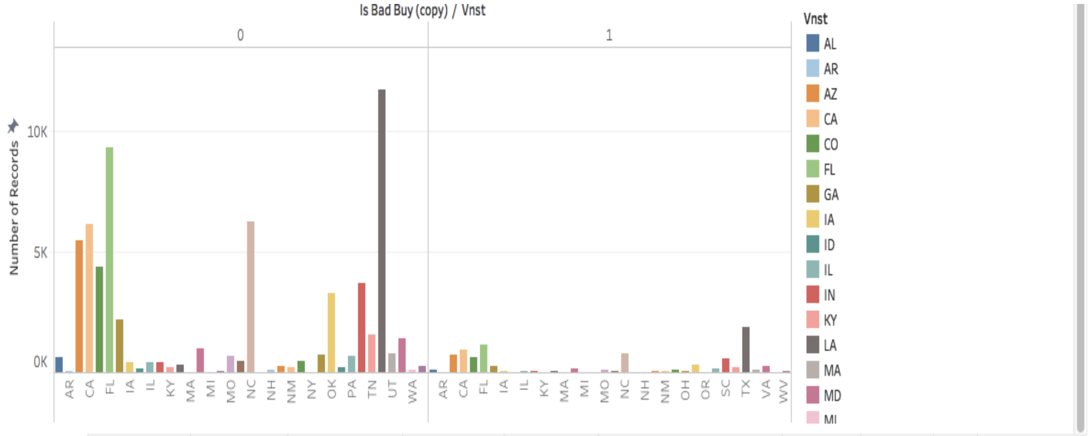
In this activity we were required to analyze the data and draw conclusions from it which would aid in solving a real world problem. We choose to differentiate a good buy from a bad buy based on a few important parameters in each state.

Data Schema:

The data set contains 32 unique features with 73,041 samples along with a labeling of 0 for good car purchases and 1 for "kicks" with the corresponding variable termed as “Is Bad Buy”. We performed our analysis on the following columns -Wheel Type, Vehicle cost, Odometer Readings, Age of the Vehicle, Make of the vehicle, Warranty Cost, IsBadBuy, Color, Size, Transmission and Vnst. We plotted each of these parameters against state.

We have come up with the following analysis for each state based on the above parameters





Inference: Using the entire analysis, we can solve a real world problem of choosing a good car for given parameters pertaining to each state.

Data Prediction

Algorithm Selection:

Before applying our prediction algorithm, we used the entire dataset to gain insight into the trend of variables and their mutual correlation with each other. This helped us understand how the variables affected each other.

For our current prediction needs, we needed to use the RandomForest classification. **Random forests** are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set, which helped us with the skewed nature of our dataset.

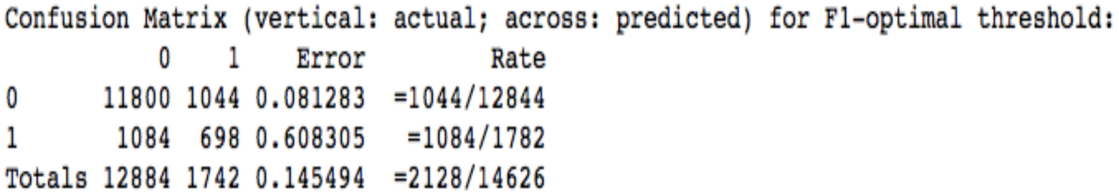
We performed our classification in R, using the package h2o package.

H2o package is an open source math engine for big data that computes parallel distributed machine learning algorithms such as generalized linear models, gradient boosting machines, random forests, and neural networks (deep learning) within various cluster environments.

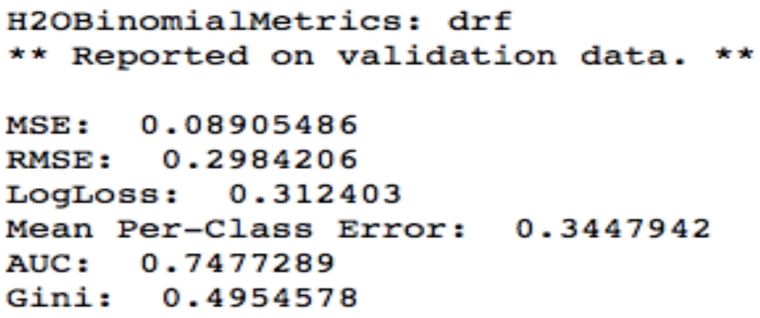
The following are the functions used:

- **h2o.randomForest** - Builds a Random Forest Model on an H2OFrame
- **predict.H2OModel** - Predict on an H2O Model

To be able to test our prediction, we split the data into training (80%) and test (20%). We used the training data to train our model and fitted this model onto the test data. We were able to achieve the following precision and rms/mse values. We also computed the confusion matrix to give us the estimates error rate.

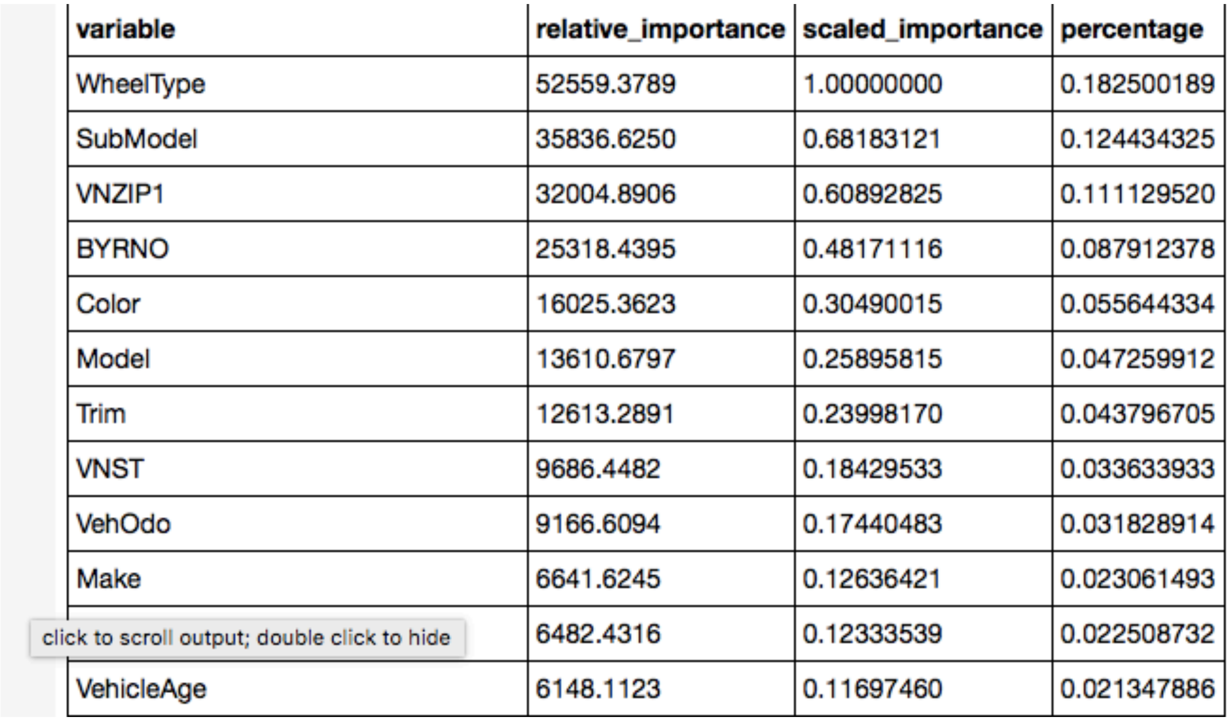


```
Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:
           0    1    Error        Rate
0      11800 1044 0.081283  =1044/12844
1       1084  698 0.608305   =1084/1782
Totals 12884 1742 0.145494  =2128/14626
```



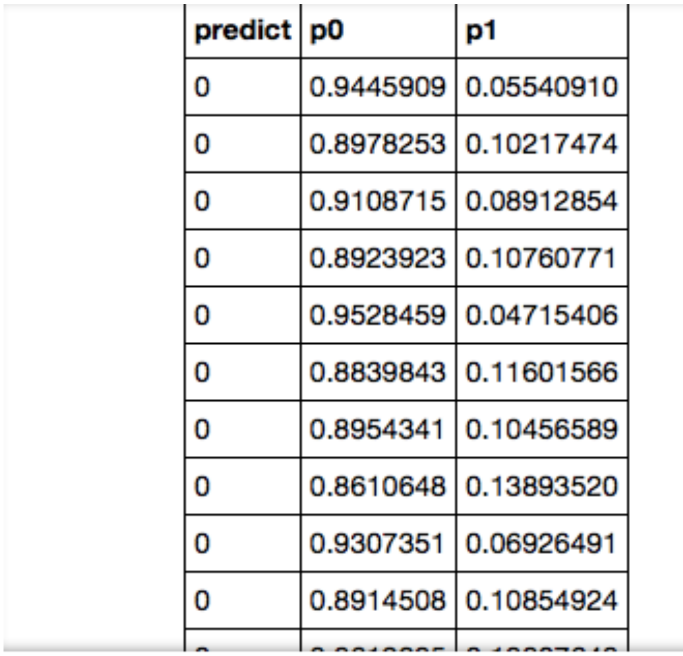
```
H2OBinomialMetrics: drf
** Reported on validation data. **

MSE:  0.08905486
RMSE:  0.2984206
LogLoss:  0.312403
Mean Per-Class Error:  0.3447942
AUC:  0.7477289
Gini:  0.4954578
```

The predicted label (0 or 1) to each test example was assigned based on a probabilistic confidence score. The label with higher probability was assigned as the examples predicted label.In addition to this, we extracted the variables which influenced the prediction the most, we observed that these variable have justified our choice of data schema



variable	relative_importance	scaled_importance	percentage
WheelType	52559.3789	1.00000000	0.182500189
SubModel	35836.6250	0.68183121	0.124434325
VNZIP1	32004.8906	0.60892825	0.111129520
BYRNO	25318.4395	0.48171116	0.087912378
Color	16025.3623	0.30490015	0.055644334
Model	13610.6797	0.25895815	0.047259912
Trim	12613.2891	0.23998170	0.043796705
VNST	9686.4482	0.18429533	0.033633933
VehOdo	9166.6094	0.17440483	0.031828914
Make	6641.6245	0.12636421	0.023061493
click to scroll output; double click to hide	6482.4316	0.12333539	0.022508732
VehicleAge	6148.1123	0.11697460	0.021347886

Variable Importance

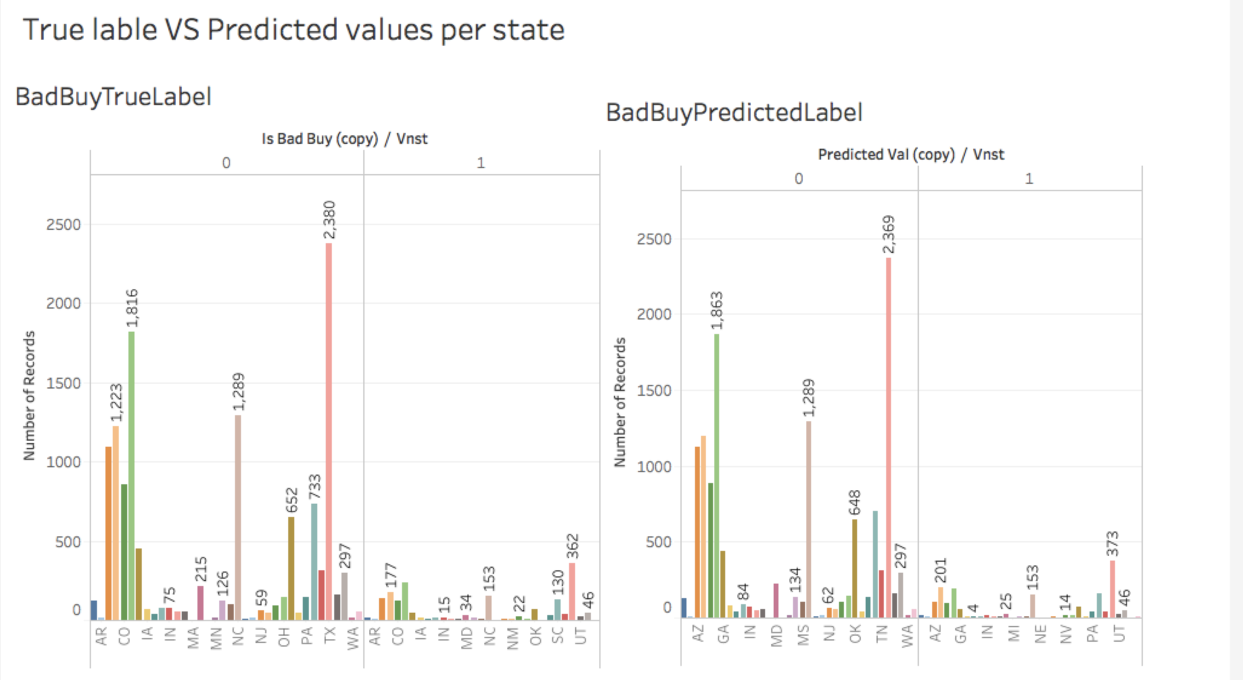


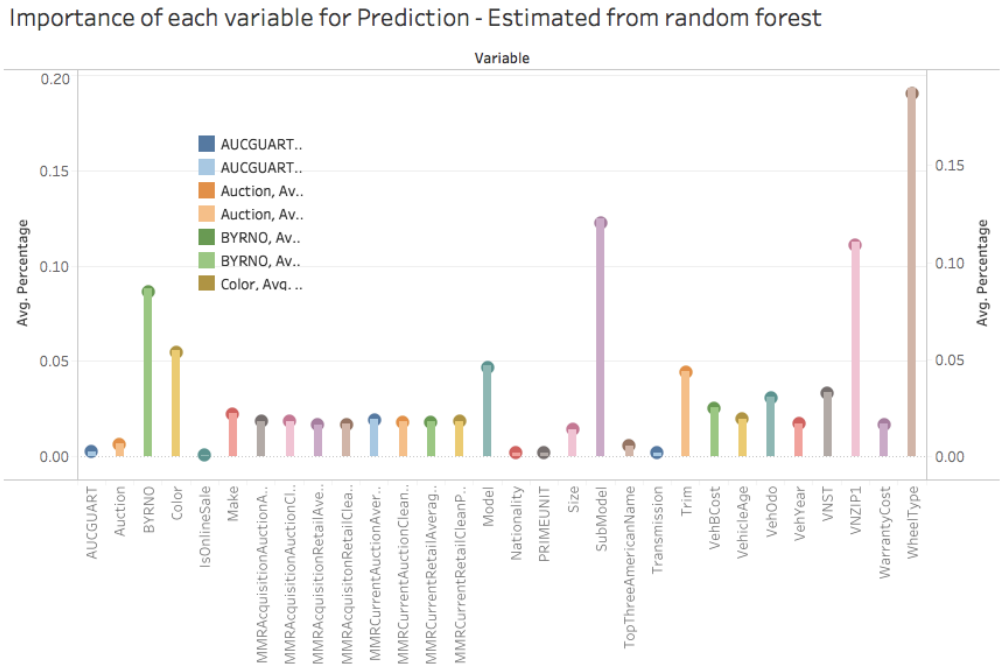
predict	p0	p1
0	0.9445909	0.05540910
0	0.8978253	0.10217474
0	0.9108715	0.08912854
0	0.8923923	0.10760771
0	0.9528459	0.04715406
0	0.8839843	0.11601566
0	0.8954341	0.10456589
0	0.8610648	0.13893520
0	0.9307351	0.06926491
0	0.8914508	0.10854924

Confidence score

Results:

We observed that random forest classification provided similar results to the true label. We then plotted the true label of the test data against the predicted label and used tableau to forecast values for the next year.





Channel: We have chosen tableau public as our channel to publish our work. Below are the links

Analysis:

https://public.tableau.com/profile/publish/Story_124/CarsthatareBadBuygivenastate#!/publish-confirm

Prediction :

https://public.tableau.com/profile/publish/Predictions_3/Predictions#!/publish-confirm