# Classification of Cars using Auction Data

Instructor: Prof. Ying Lin

Team: Bhavish Kumar, Sai Praharsha Devalla, Tejas Patil

## Problem Statement

- Auto dealers who purchase used cars at an auto auction face the risk of buying a bad car with a lot of issues, because of which they will not be able to re sell it to their customers.
- These bad cars can be a huge financial burden on the car dealers because of all the transportation, throw-away and repair costs that they must bear with.
- By predicting if a car is going to be a bad buy or not, we can help the dealers make wise and informed decisions on the cars that they need to purchase which can in turn help them minimize their incurred loss and maximize profits.

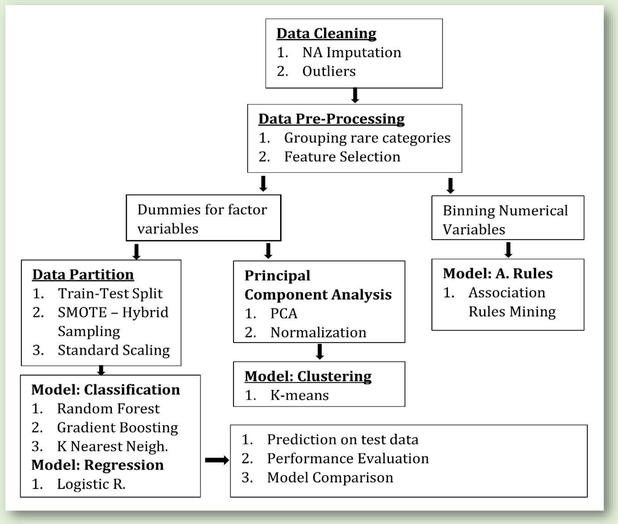
## Project Objective

- The objective of the project is to help the car dealers accurately predict if a purchased car is going to be a good buy or a bad buy.
- The predictions made will help the car dealers take informed decisions on the cars that they should go ahead and purchase and on the cars that they should refrain from purchasing.
- We will also try to identify the most important attributes that will help us predict if a purchased car is going to be a good buy or a bad buy.

## Dataset Description

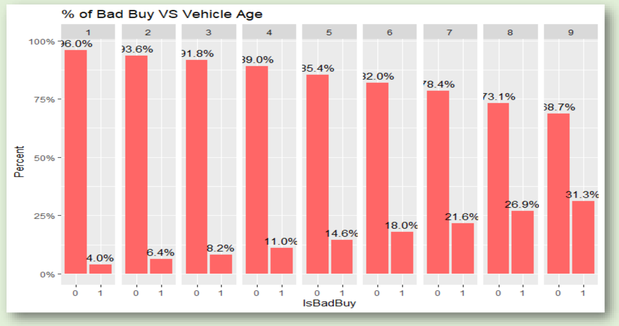
- The Carvana cars dataset is a Kaggle competition dataset with 34 assessment parameters, one of which is the target variable.
- The binary target attribute have 'good buys' and 'bad buys' approximately in the ratio 7:1 respectively. The large disparity between the counts of 2 classes in the target variable makes our dataset imbalanced.
- Data variables describe various specifications of auctioned car and the acquisition and current price of that car during auction and retail sales.

## Data Cleaning and Pre-Processing




*Project Flowchart*

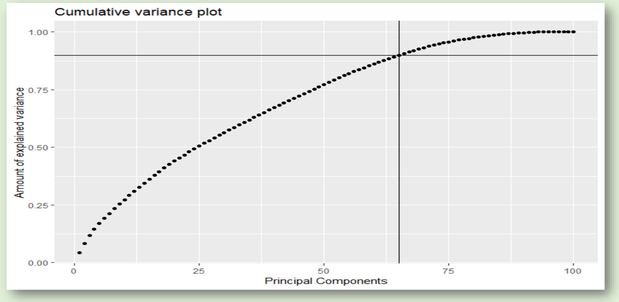
## Data Exploration




## Pre-Processing Techniques

| Techniques | Purpose |
|---|---|
| Chi-Square Test of Independence | Feature Selection: Determining relation between independent categorical variables and target variable. |
| One Hot Encoding | Processing and coding categorical variables into dichotomous variables. |
| Principal Component Analysis | Reducing the dimensions of the data but capturing most of the variance. |
| SMOTE: Sampling | Performing combination of over sampling and under sampling on target variable. |
| Standard Scaling | Scaling the range of all variables so that each feature contributes to the final distance. |
| Normalization | Normalizing the scores of principal components. |

## Principal Component Analysis




## Data Mining Algorithms

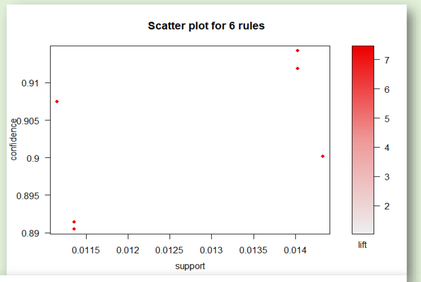
| Type | Algorithm | Evaluation Parameter |
|---|---|---|
| Classification | Random Forest | Accuracy, Recall |
| Classification | Gradient Boosting (gbm) | Accuracy, Recall |
| Classification | K-Nearest Neighbor | Accuracy, Recall |
| Regression | Logistic Regression (glm) | Accuracy, Recall |
| Clustering | K-means (Hartigan-Wong) | Sum of Squared errors |
| Association Rules | Apriori | Support, Confidence, Lift |

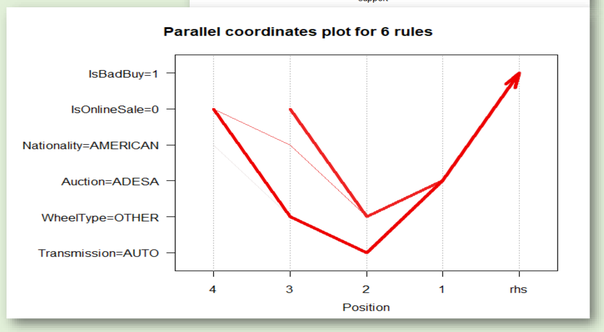
## Clustering


*Elbow curve and Cluster representation*

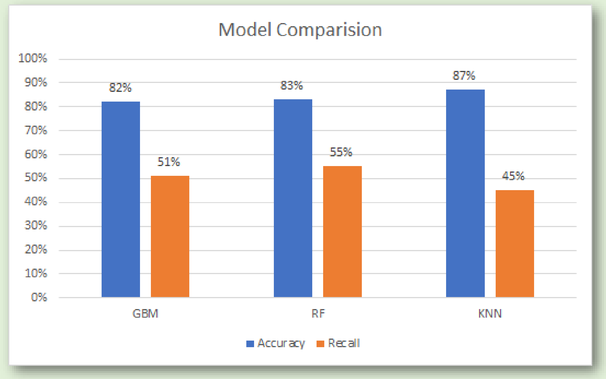
## Association Rules

*Scatter plot and paracord plot representation of best rules extracted Bad purchases*

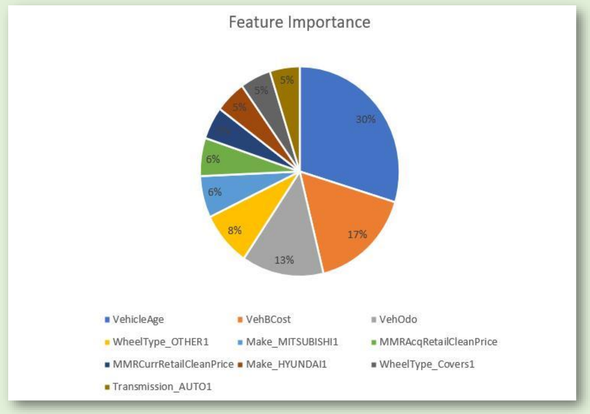



## Classification




## Feature Importance




## Conclusion

- When a vehicle dealer goes to the car auction, biding the money on the right car is very important. Our application can help the dealer in reducing the risk of purchasing a car that turns out to be a bad buy and then concentrate on cars that are going to be a successful purchase. The application also reveals how the car's features lead to a bad buy for that vehicle.

## Future Scope

- As we ran algorithms with different hyperparameters, due to imbalance in the dataset we found a trade off between accuracy and recall. The project work can be continued further using an enhanced algorithm to deal with class imbalance that will help improve both accuracy and recall values.

## References

https://www.kaggle.com/c/DontGetKicked

https://medium.com/airbnb-engineering/confidence-splitting-criterions-can-improve-precision-and-recall-in-random-forest-classifiers-ad2d4ba696a4

https://towardsdatascience.com/imbalanced-data-in-classification-general-solution-case-study-169f2e18b017