KAGGLE COMPETITION - DON'T GET KICKED!

Predict if a car purchased at an auction is a bad buy

CMPE 239 Web and Data Mining project

Submitted to Prof. Magdalini Eirinaki

Haritha Peyyeti

Yi (Christy) Chou

MOTIVATION

- Biggest challenge at vehicle auctions
- Risk of buying cars with serious issues
- Bad cars called as "kicks" or "lemons"
  - Tampered odometers
  - Mechanical issues
- Can a model identify which cars have a higher risk of being a *kick?*

OBJECTIVE

- **Goal**: Develop a prediction model that dealerships can utilize when buying used cars at auctions.

- Minimize the risk of buying bad cars

- Reduce the losses incurred

- Provide their customers with best inventory selection.

ALGORITHMS CONSIDERED

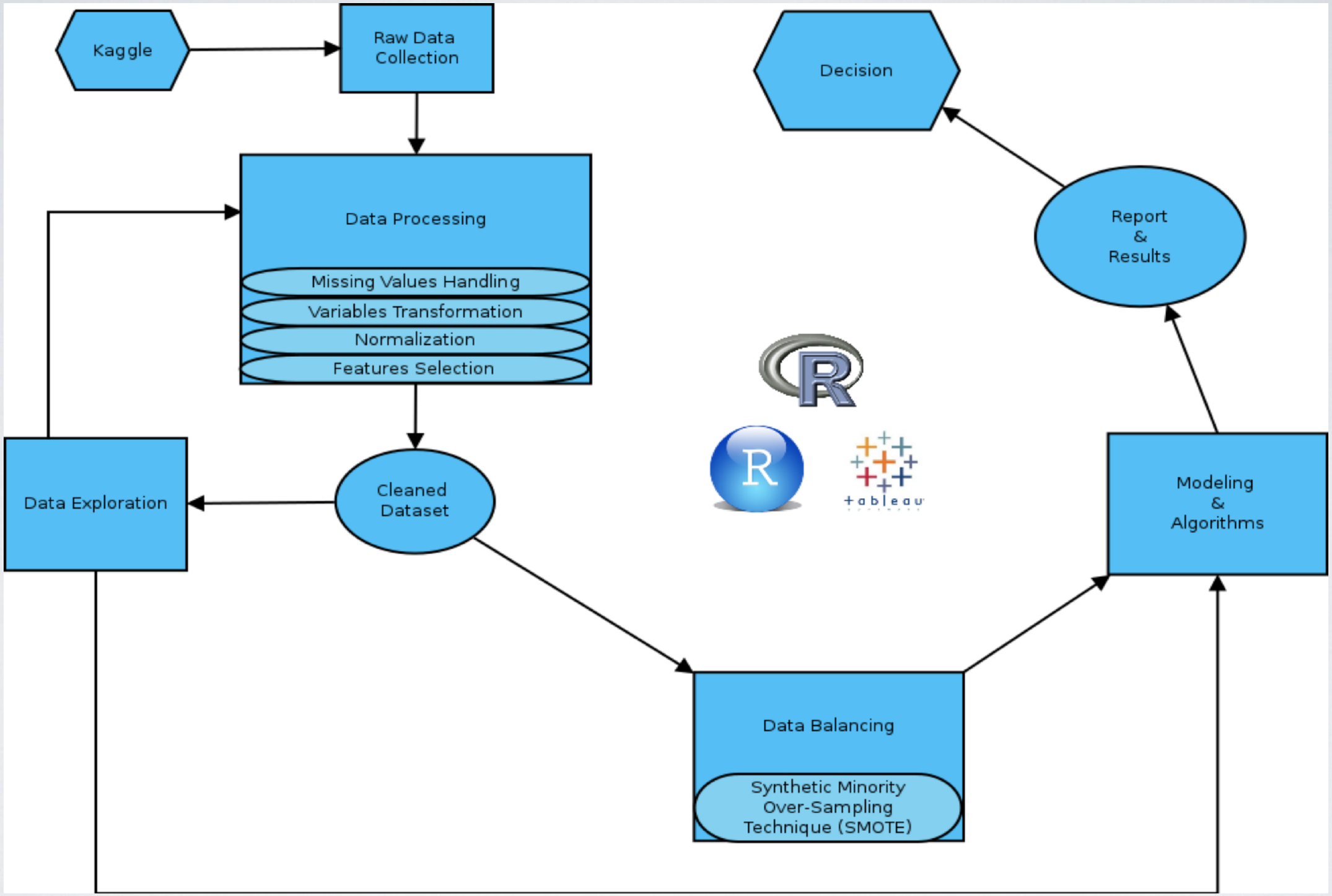
- Logistic Regression
- Naïve Bayes
- Support Vector machines
- Logitboost
- Classification and Regression Tree (CART)
- Adaptive Boosting (AdaBoost)
- Random Forest

TECHNOLOGIES USED

- Tableau
  - Data exploration
- R : Programming language
  - Pros:
    - Abundance of packages
    - Ease of data visualization and exploration
    - Analyze statistical aspects
  - Cons:
    - Memory issues! ?
    - How to tackle?
- Amazon EC2 micro instance

Package Name	Algorithm / Function
pROC	Visualization and analysis of ROC curves
FSelector	Attributes Importance Calculation
caret	***C***lassification ***A***nd ***RE***gression ***T***raining for predictive models
caTools	For utility functions
DMwR	Synthetic minority over-sampling technique (SMOTE)
rpart	Classification and Regression Tree
e1071	Naïve Bayes
adabag	AdaBoosting
ROCR	Visualization of classifier performance measures
rminer	Classification and Regression algorithms
ggplot2	Data Visualization
randomForest	Random Forest

Section 2: System Design



USE CASE


'Kick' car predictor
Vehicle Make
Select Brand the list
Vehicle Age
Enter years since manufacture
Odometer Reading
Enter mileage
Wheel type
Select wheel type
Market Prices
Import
Submit


'Kick' car predictor
Vehicle Make
Select Brand the list
Vehicle Age
Enter years since manufacture
Probability of a 'kick' is 0.895
Market Prices
Import

DATA DESCRIPTION

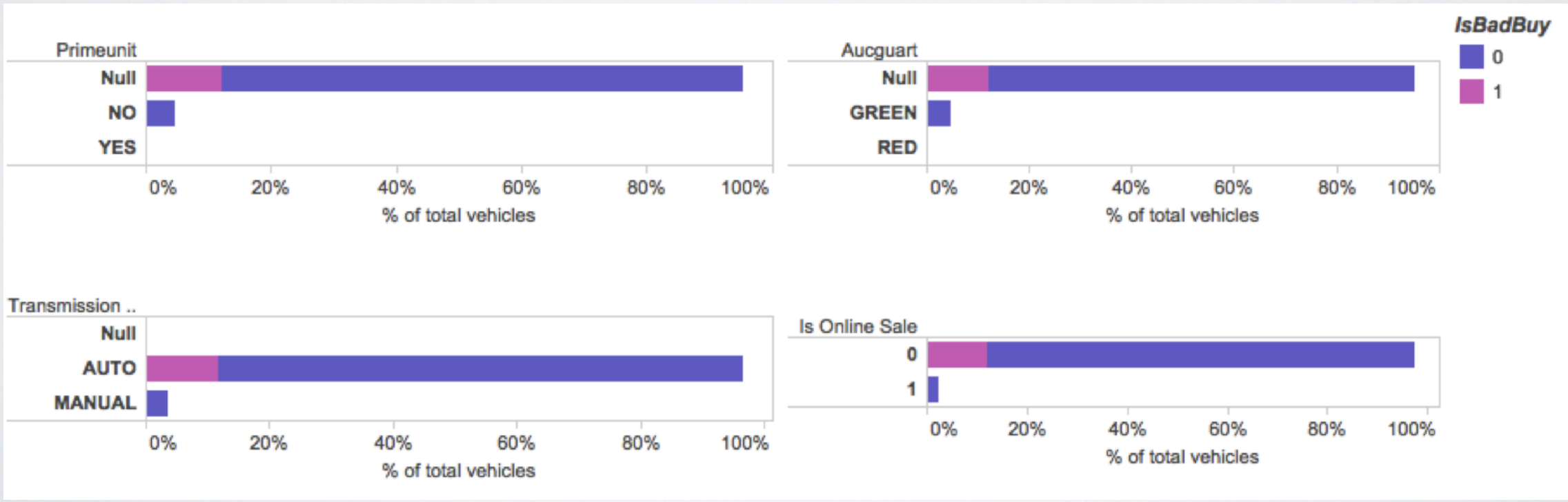
- Training Data: 32 features and 72,983 samples
- Predictors include:
- Specifications of the vehicle - make,model,age
- Market Prices
- Purchase specific details

DATA PREPROCESSING

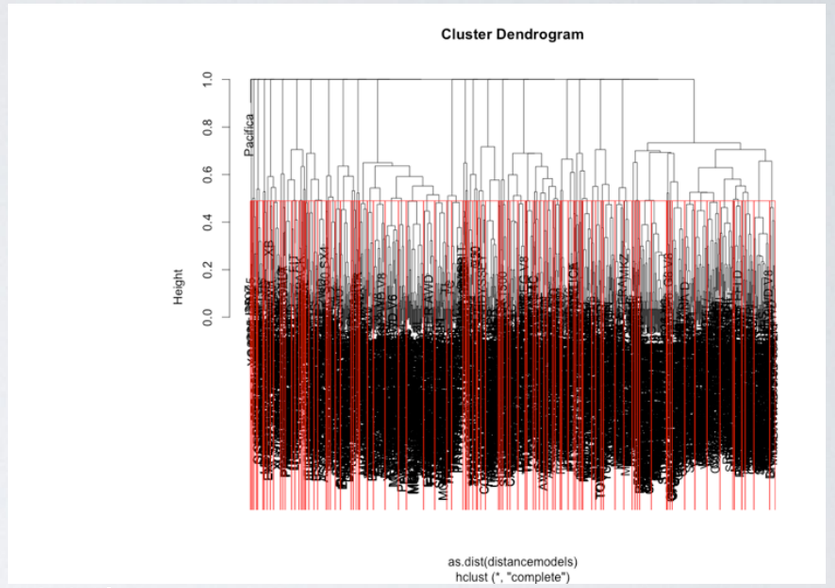
- **Missing Value Treatment:**
  - Numeric variables: replace with median value
  - Categorical variable, missing values are grouped to a new category.
- **Redundant Variables: Causes Multicollinearity**
  - Wheel type
  - Wheel type ID
- **Poor Quality Variables: Adds no value**
  - More than 50% of the values is null

DATA PREPROCESSING

- **Near Zero Variance Variables**: Adds no value
  - Unique values for over 95% of total sample



DATA PREPROCESSING



- **Variables Transformation**:
  - Stand-alone absolute values add little value
  - Transform absolute market prices to relative prices
    - Difference between retail and auction price adjusted by the vehicle age.
- **Variables with Large Number of Text Factors**:
  - Vehicle model: 1130 levels Vehicle sub-model: 934 levels
  - Jaro-Winkler distance method to approximate string factors for reducing factor levels

DATA PREPROCESSING

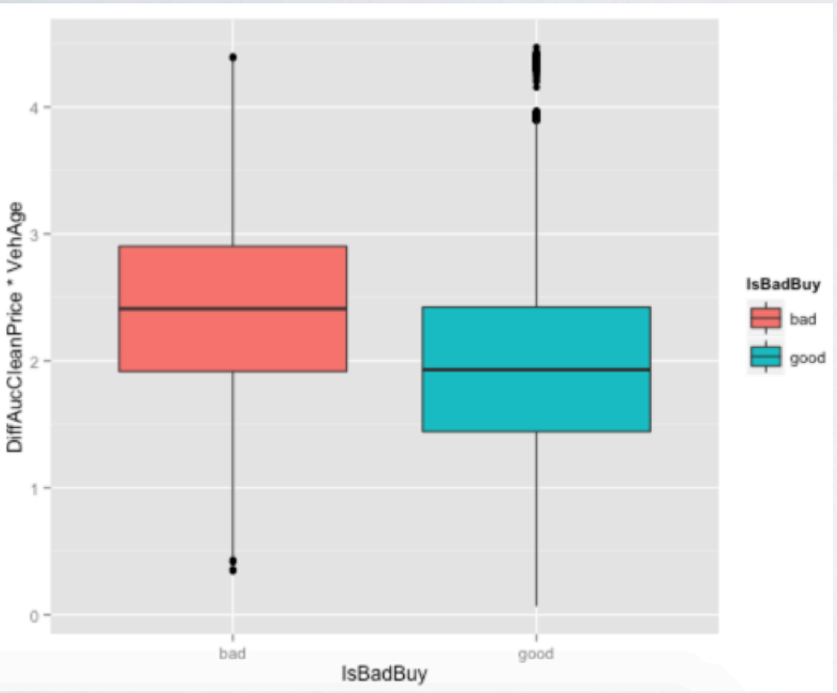
- **Data Balancing**:

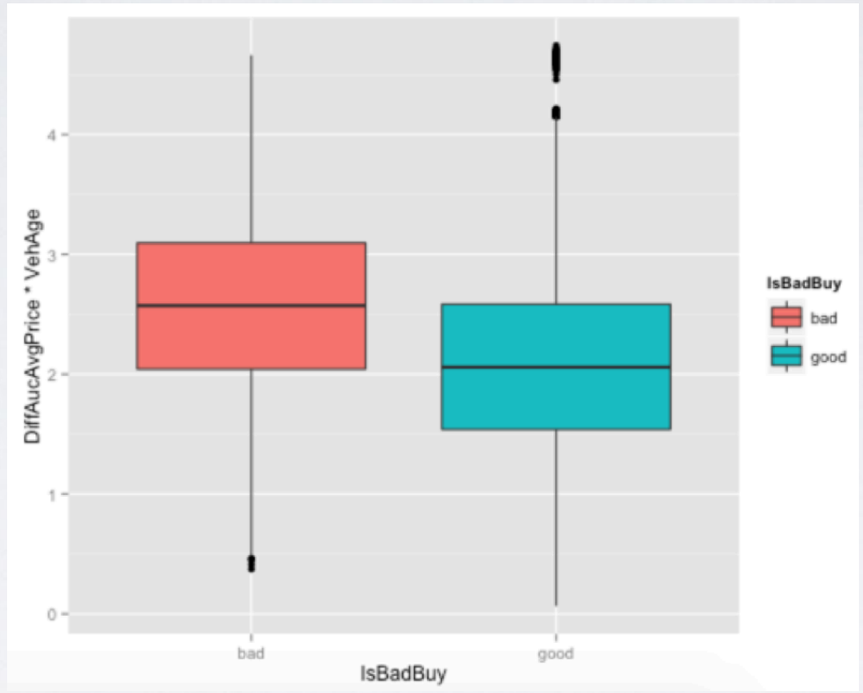
  - Original data is unbalanced with 87.7% good instances and 12.3% bad instances.

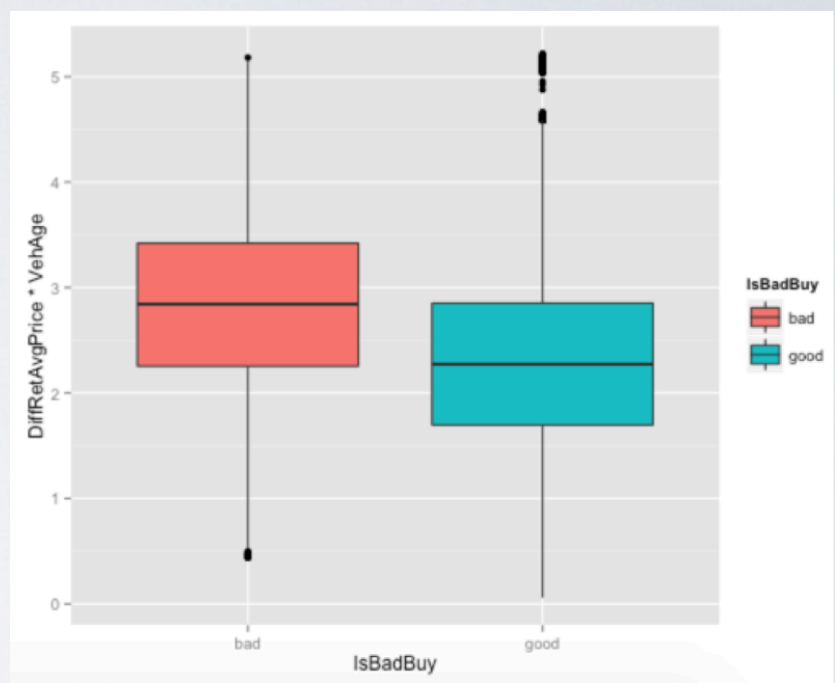
  - **S**ynthetic **M**inority **O**ver-sampling **Te**chnique (SMOTE) Algorithm: over-sampling 5 times more bad instances and under-sampling 1.2 times less good instances.

  - Balanced data with 50% good instances and 50% bad instances.

DATA VISUALIZATION



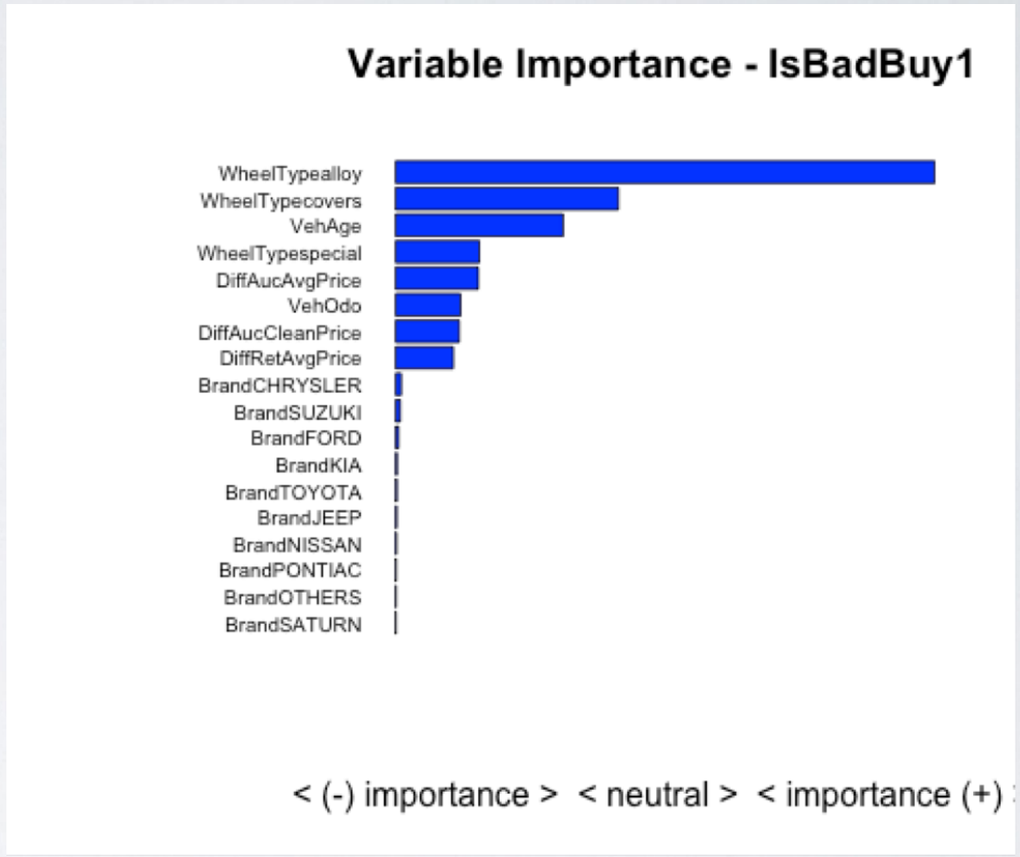




- Difference of auction or retail price for bad cars are generally higher than for good cars

METHODOLOGY

- **Feature Selection**
  - Chi-squred filter
  - Information gain
  - Step-wise logistic regression method



- **Data Leak**
  - Excluded attributes specific to the purchase
  - Unattainable at the time of prediction
  - E.g. VehBcost (price paid for the vehicle) ,BYRNO (Buyer number)

MODEL BUILDING

- **Unbalanced data**
    - Training set: 51,088 samples
    - Test set: 21,895 samples
    - Bad cars prevalence :12%
- **Balanced data**
    - Training set: 107,712 samples
    - Test set: 21,895 samples
    - Bad cars prevalence :50%
- Models built using balanced and unbalanced data
- 5 fold cross-validation technique
- Use the best model after adjustments to make predictions for test set.

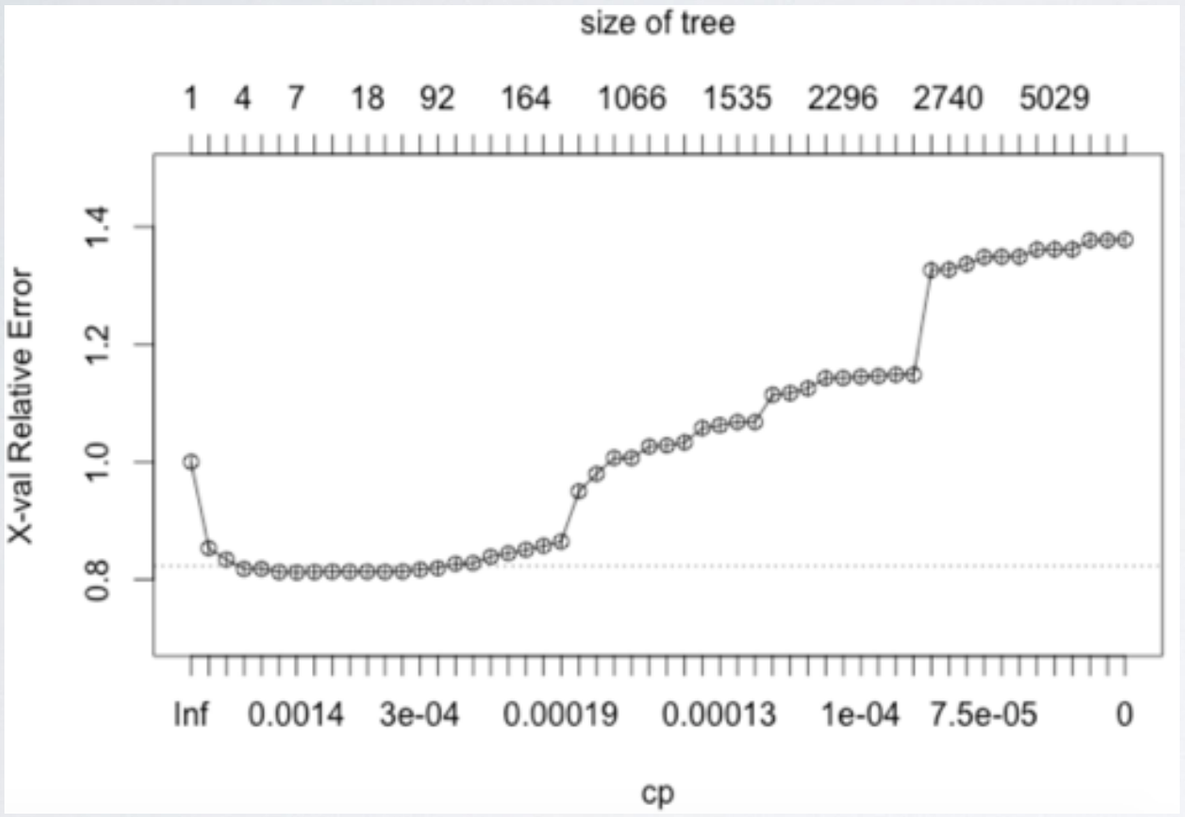
PERFORMANCE EVALUATION MEASURES

- Accuracy?
  - Underrepresentation of bad cars
- Precision
- Recall
- F-measure
- Area Under Curve (AUC)

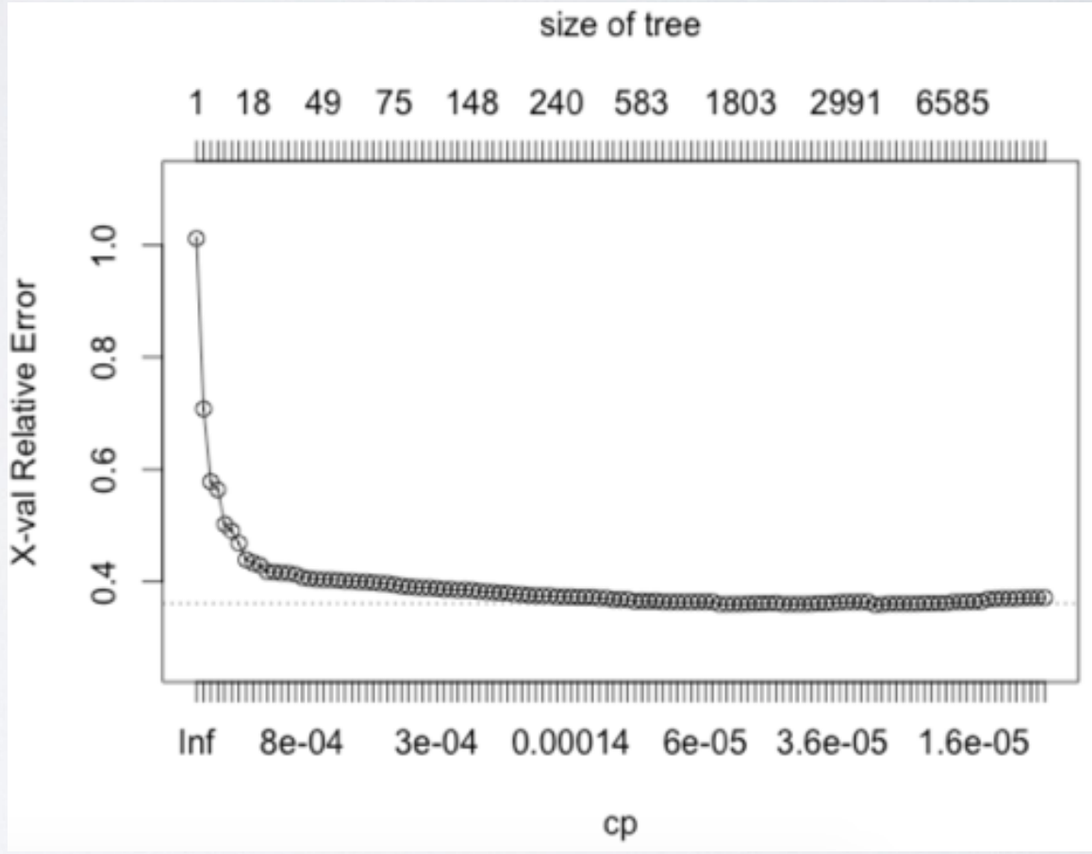
EVALUATION OF ALGORITHMS

Classification and Regression Trees (CART)

Using Unbalanced training set



Using Balanced training set



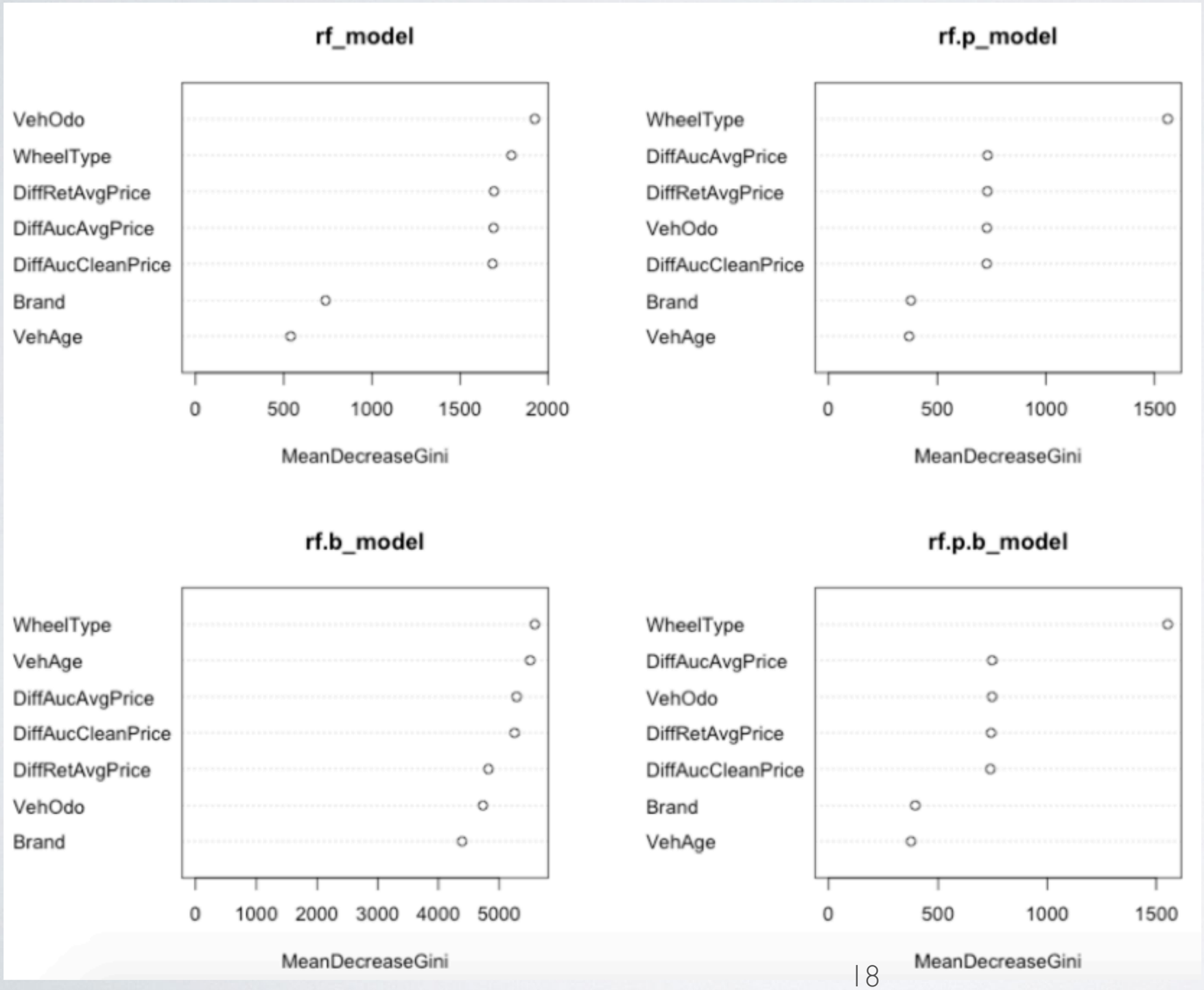
For pruned trees:

AUC: 0.601
F-measure: 0.333

AUC: 0.612
F-measure: 0.306

EVALUATION OF ALGORITHMS

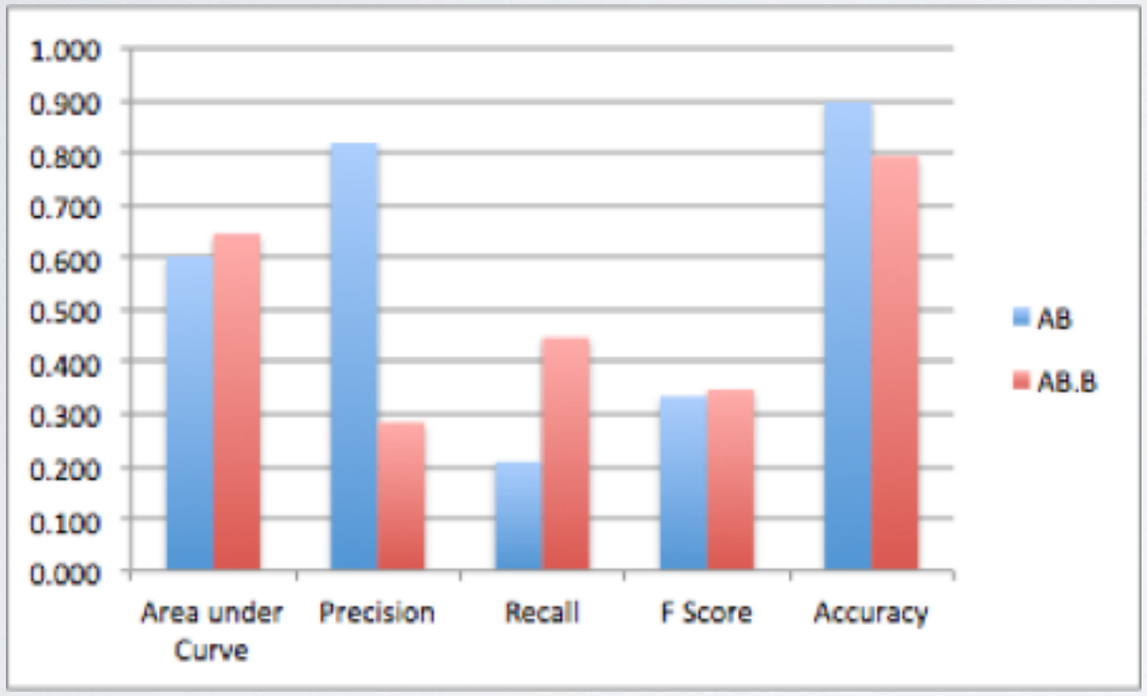
Random Forest (RF)



- Tuning does not help improve the model
- AUC = 0.8569 and F-measure is 0.661 for un-tuned RF

EVALUATION OF ALGORITHMS

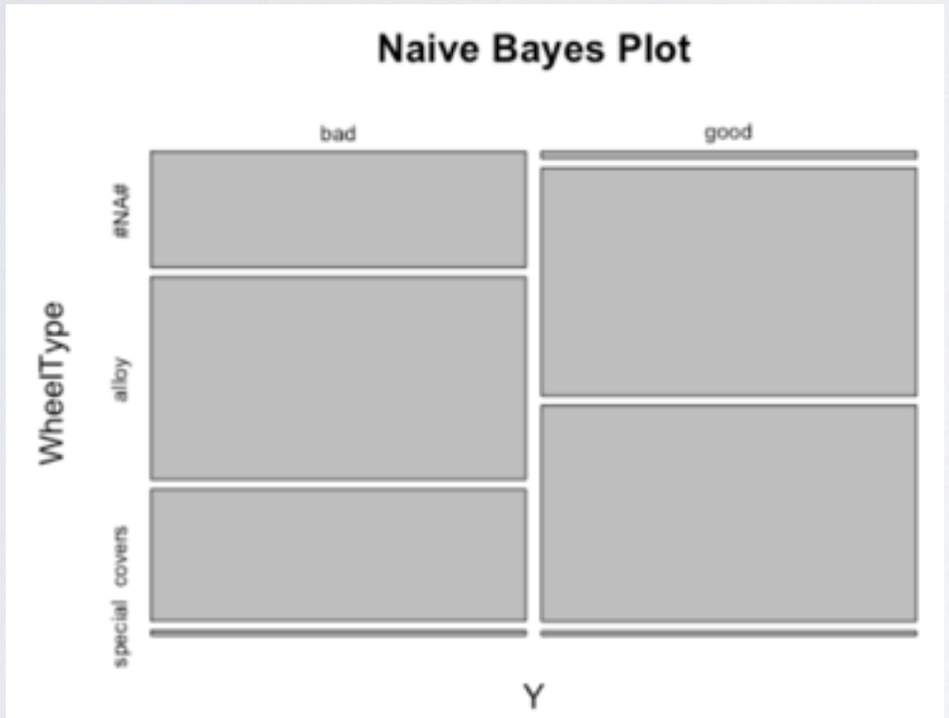
Adaptive Boosting (Adaboost)



- For unbalanced training set, AUC = 0.6015 and F-measure is 0.334
- For balanced training set, AUC = 0.6445 and F-measure is 0.347

EVALUATION OF ALGORITHMS

Naïve Bayes (NB)

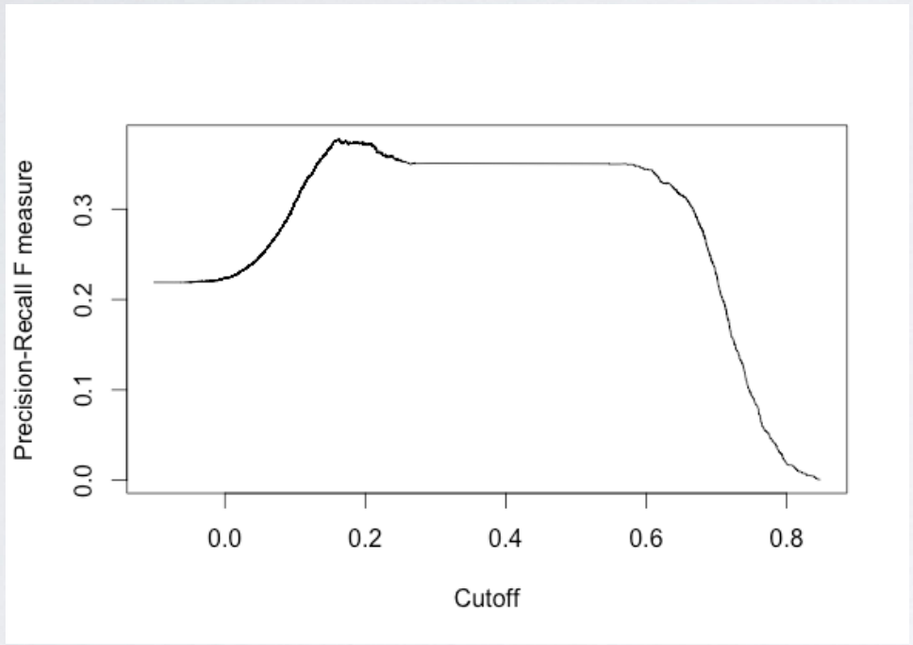


- Mosaic plot of Naive Bayes Algorithm
- Numerical probability close to zero for all classes for many observations
- Tuned Laplace Smoothing parameter, generating same problems

EVALUATION OF ALGORITHMS

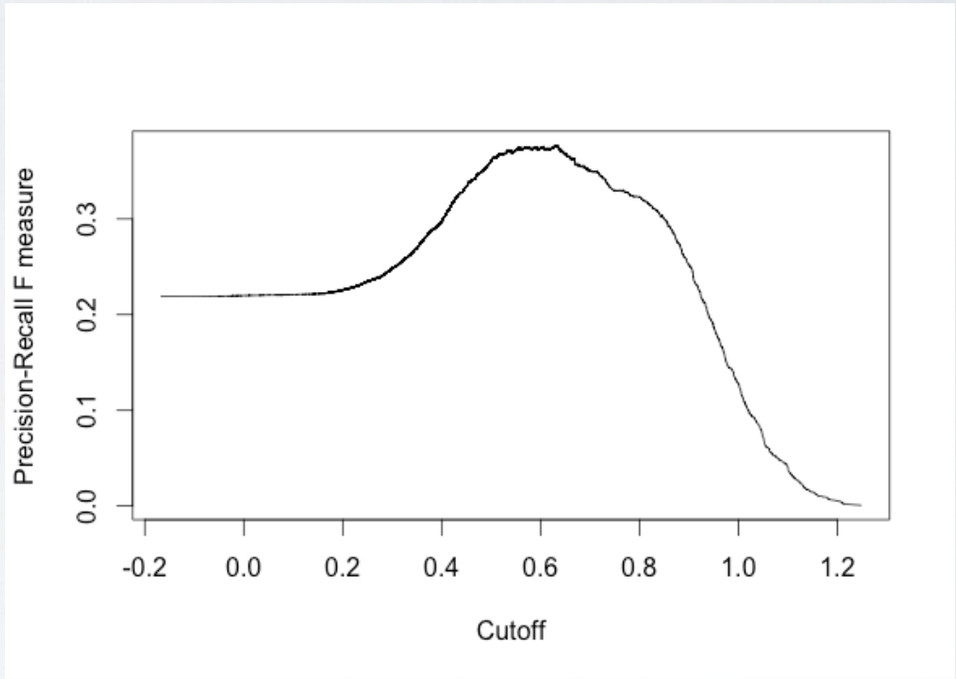
Logistic Regression

Using Unbalanced training set



AUC: 0.736
F-measure: 0.372

Using Balanced training set

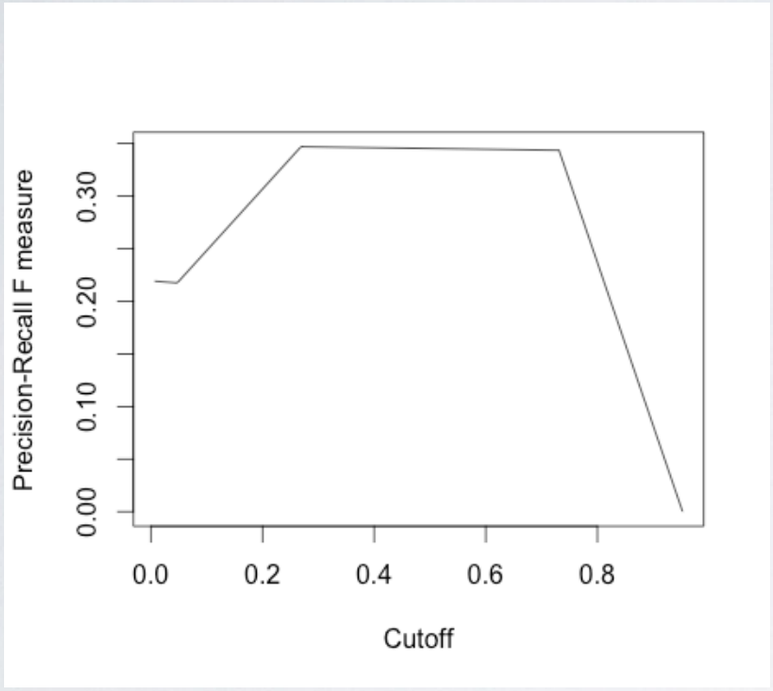


AUC: 0.734
F-measure: 0.374

EVALUATION OF ALGORITHMS

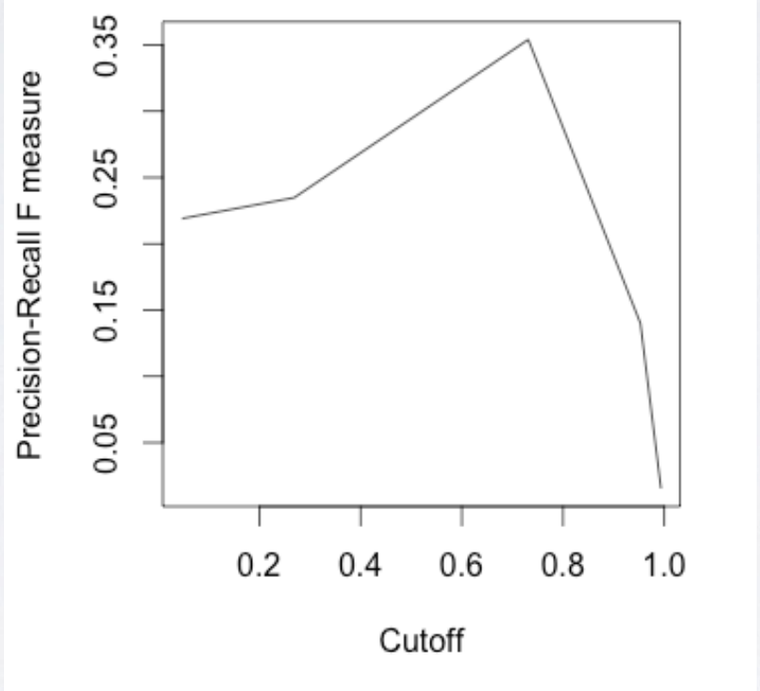
Logitboost

Using Unbalanced training set



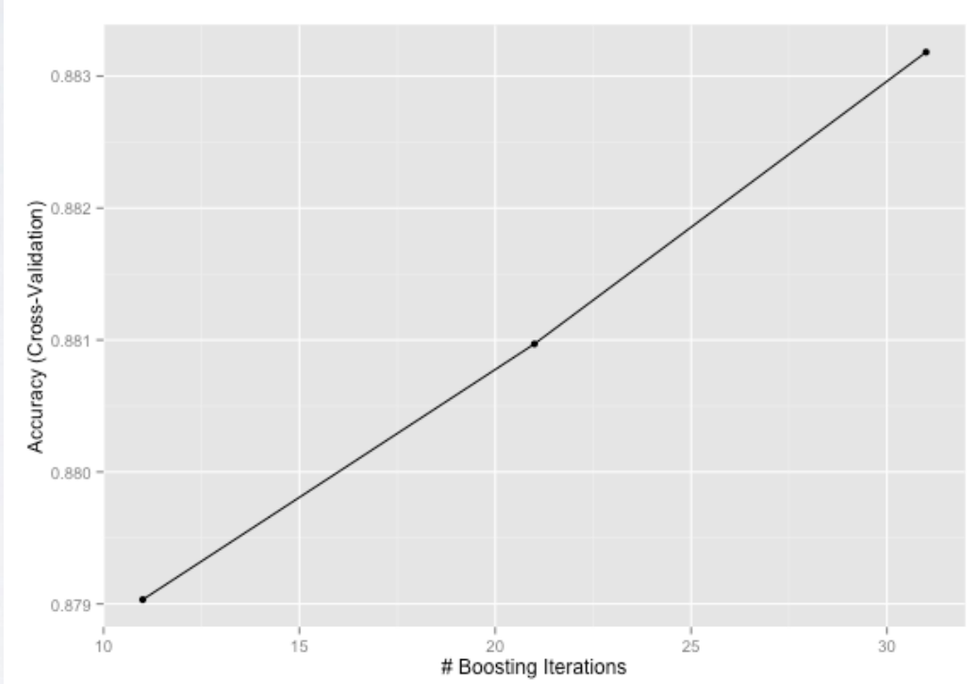
AUC: 0.734
F-measure: 0.344

Using Balanced training set



AUC: 0.745
F-measure: 0.140

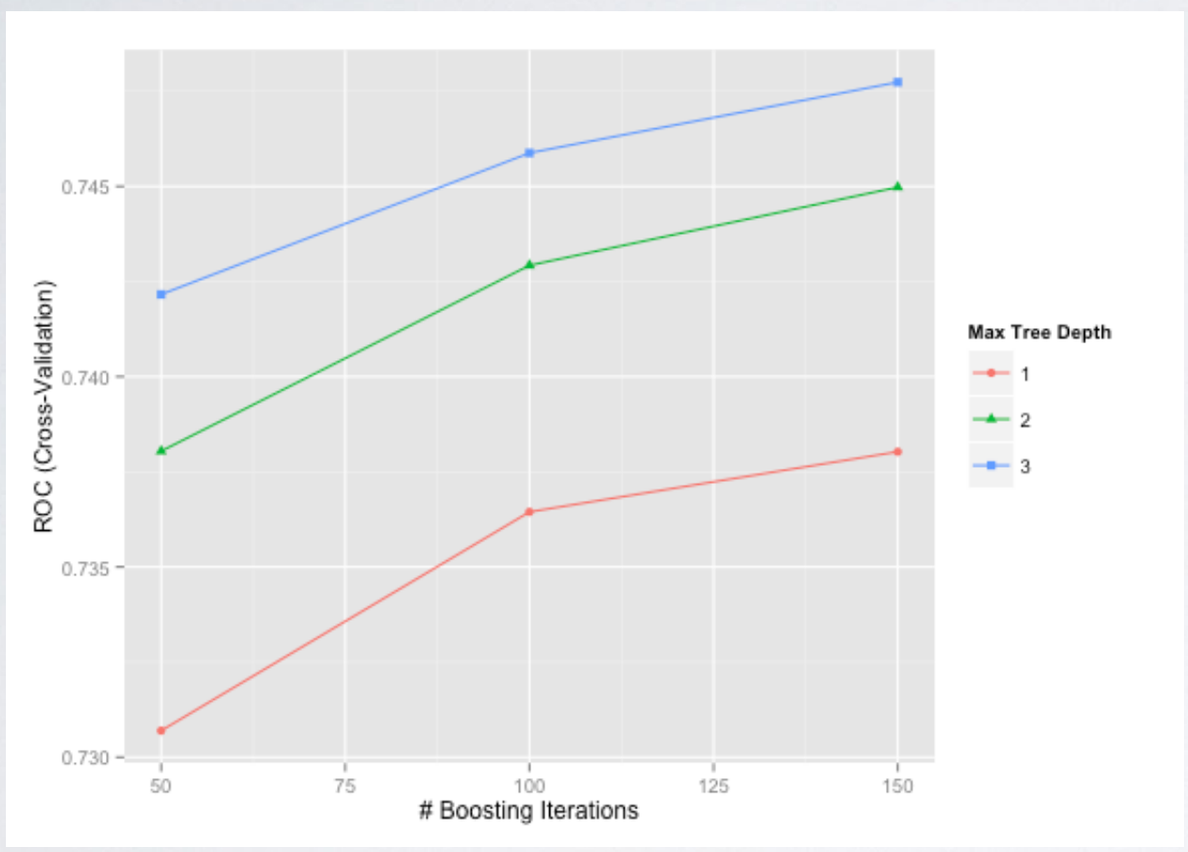
Accuracy plot



EVALUATION OF ALGORITHMS

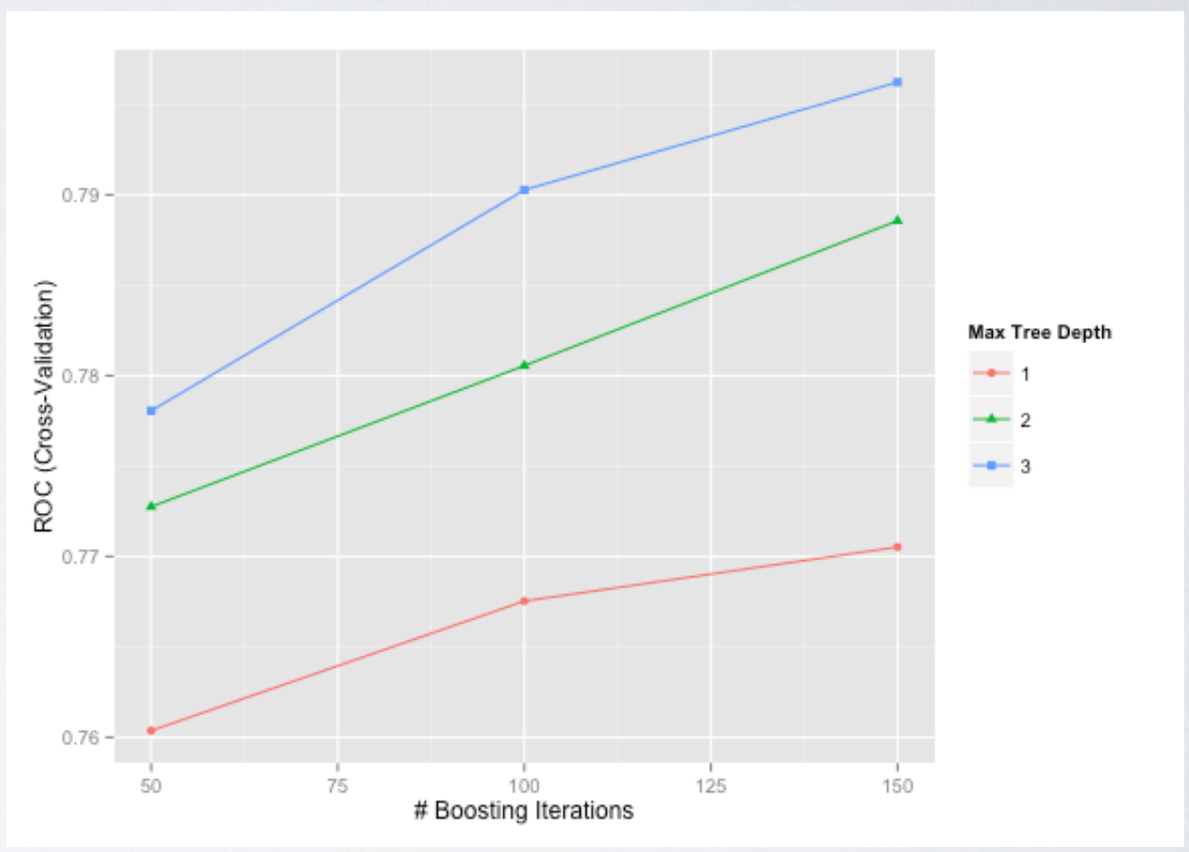
Gradient Boosting Method (GBM)

Using Unbalanced training set



AUC: 0.744
F-measure: 0.372

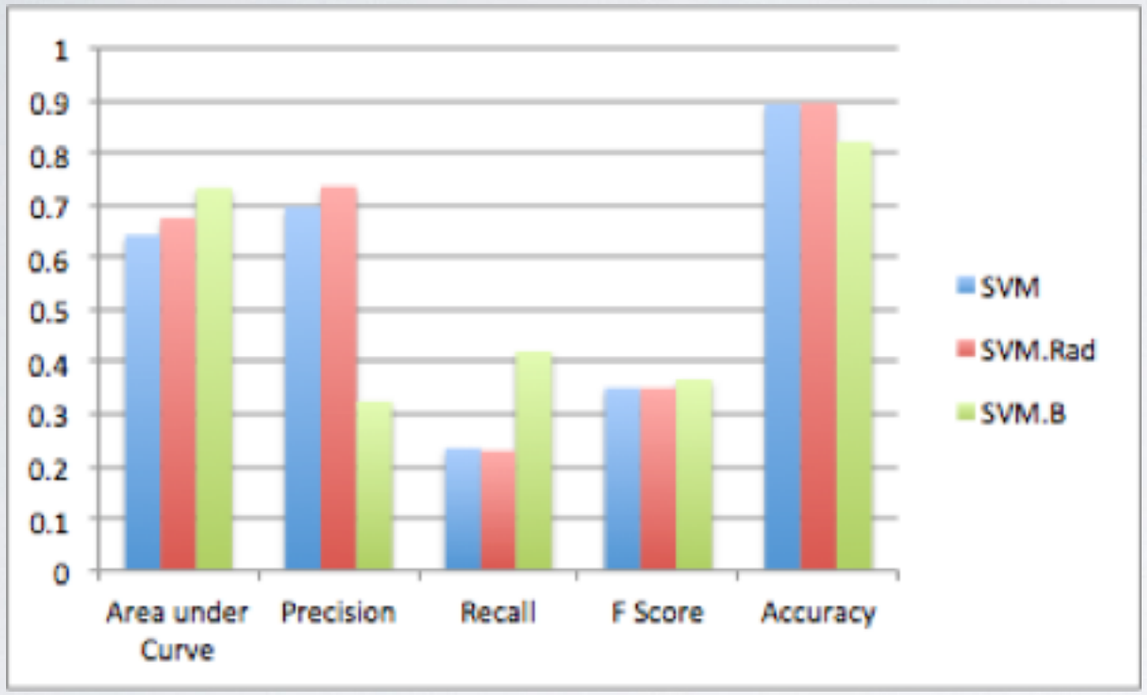
Using Balanced training set



AUC: 0.745
F-measure: 0.372

EVALUATION OF ALGORITHMS

Support Vector Machine (SVM)



Original SVM

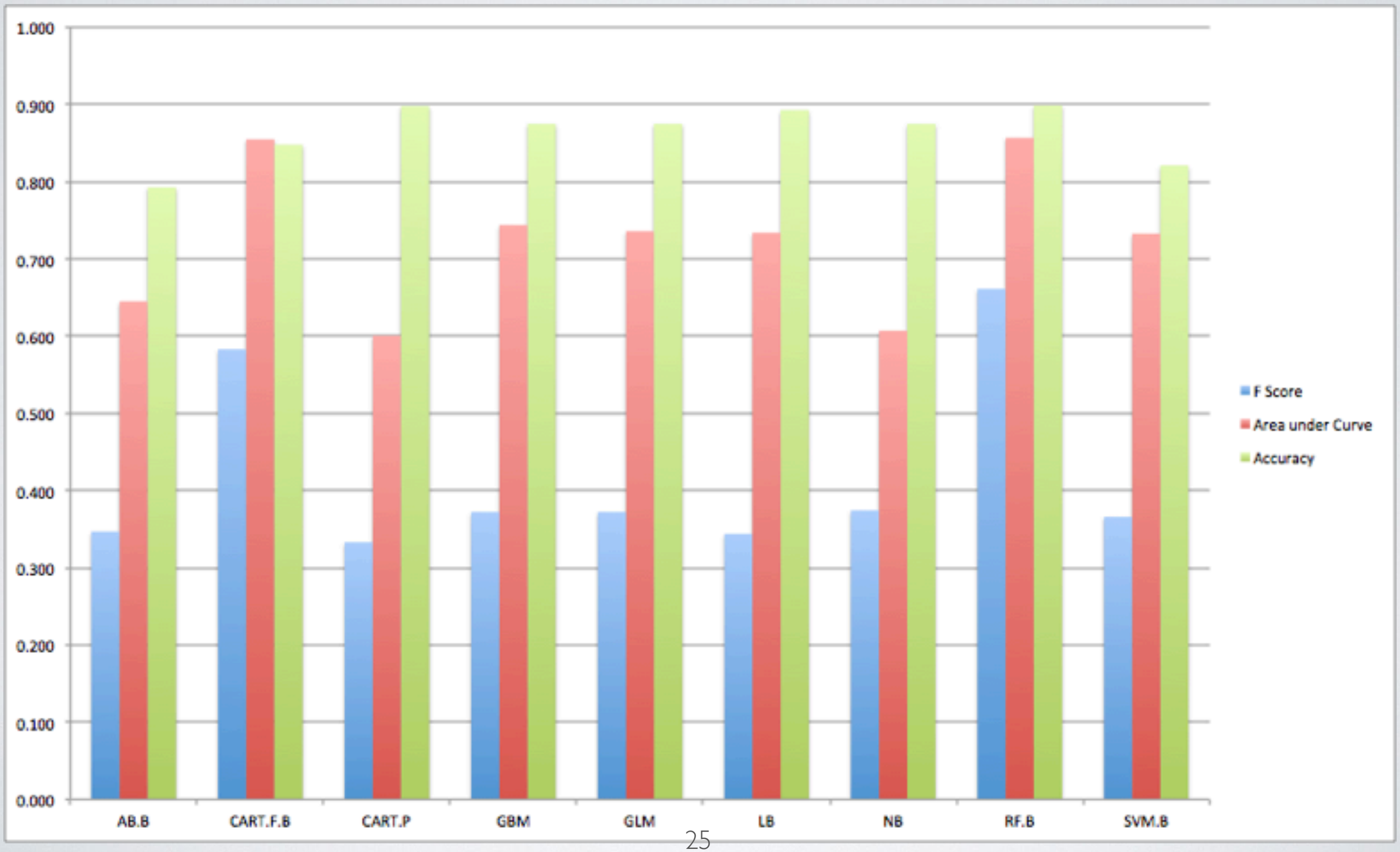
- For unbalanced training set, AUC = 0.6418 and F-measure is 0.350
- For balanced training set, AUC = 0.7329 and F-measure is 0.366

SVM Radian

- AUC = 0.6749 and F-measure is 0.350

EVALUATION OF ALGORITHMS

Comparison of Algorithms



DISCUSSION

- **Things worked well:**

- Effectively handle features for reducing computation time.

- Experiment with balanced data

- **Difficulties faced:**

- Computationally difficulty for Support Vector Machine, AdaBoost, and Random Forest with repeated cross validations for k-folds algorithm.

CONCLUSION

- Random Forests and Support Vector Machine build better predictive models for this data set.

- WheelType and VehOdo are the most important variables to check when purchasing cars.