The Data Challenge: Identifying Bad Auction Purchases

nerdwallet

Thilak Dasarathan

Contents

Goals

Classify **lemons** from the good auction purchases for the auto company

Data Assessment

Data was cleaned, assessed and imputed to be ready for further analysis

Profile

Attributes of all the auto purchases were profiled to understand it’s relationship of the purchase being a lemon

Predict

ML/Statistical models were trained and validated to predict the bad purchases in auction for auto company

Insights

Business Insights were extracted from the classifiers to ensure more educated purchases to avoid lemons

Roadmap to Achieve our Goal

The process below were used to evaluate the data; each and every variable was studied for their statistical value in both univariate and multivariate environment

Univariate Analysis

Data Processing		Feature Engineering		
Ingest	**Processing**	**Impute**	**Assess**	**Go/No Go**
Read-in data, merged JSON with CSV, De-Duped	Translated, created new variables to assess usability	Impute missing values with appropriate information for better use of data	Measure statistical value and relationship with lemon purchases	Determine strategic and statistical value

Multi-variate Analysis (Model Building)

Ingest	Train	Assess	Validate	Score
Ingest the variables that passed round 1 and use learnings towards multivariate analysis	Train the model with the variables studied and refine it in a multivariate setting	Understand the outputs from the model and compare it against the univariate analysis performed	Validate the model on the validation sample and refine the model to reduce variance & bias in the prediction	Score the test dataset with the best trained model with the balanced variance & bias in prediction

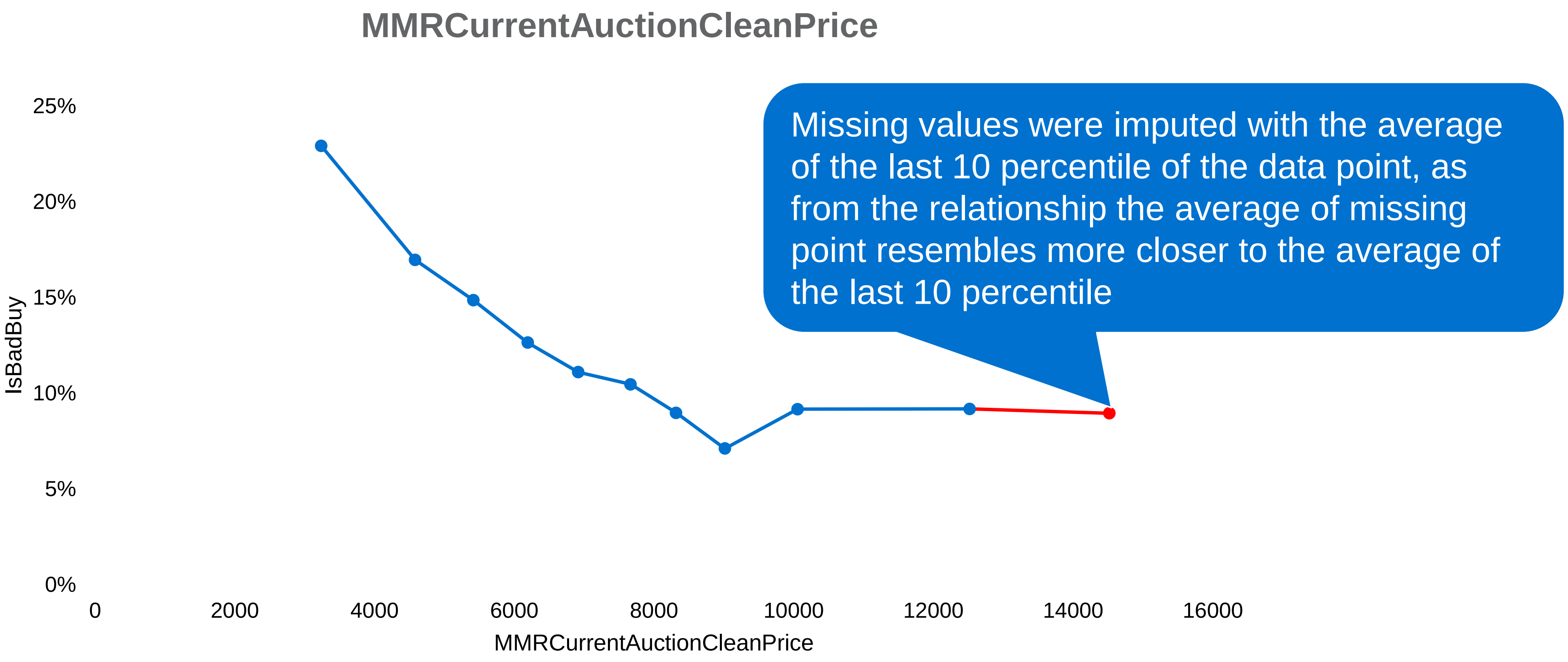
Data Processing

The data provided has been ingested and cleaned for analysis

- Identified & De-Duped the duplicate rows in the train dataset
- Ingested the JSON file and mapped it to Train/Test dataset
- Additional variables were created while cleaning the dataset
  - Dummy variables (Binary Variables) are created for categorical variables like Size, Wheel Type, Transmission, etc..
  - Ratio of the Paid Cost (VehBCost) with the other Acquisition Price (MMRAcquisitionAuctionAveragePrice,etc..)
  - Vehicle Age – Difference between the vehicle purchase year and vehicle manufacturer’s year

Feature Engineering: Impute Missing Values

The purchases with missing data for certain price related variables were imputed based on their relationship with the IsBadBuy variable (Dependent Variable)



All the price related variables were evaluated one at a time and imputed missing's with more informed values

Feature Engineering : Univariate Assessment

Conducted several statistical tests to understand and measure the relationship and predictive power of the variables

In priority order, the below tests were conducted for all the variables

- **Association of the independent variables to have a strong relationship with IsBadBuy (Dependent Variable)**
  - Measured by calculating correlation analysis and conducting Chi-SQ
  - Higher the Chi-SQ or Correlation Coefficient stronger the relationship of a variable with dependent variable

- **Variance of the dependent variable explained by the purchase related variables/independent variables**
  - This metric was conducted only for continuous variables
  - Measured using F-Test
  - Higher the F value, higher is the variance explained by a variable

- **Simplicity of the variables relationship with IsBadBuy**
  - This metric was conducted only for continuous variables
  - Measured by calculating R-Square of the variables with the dependent variable
  - Higher the R-Square, stronger and simpler the relationship

- **Spread explained for the dependent variable**
  - Measured by calculating the Max – Min for the variables
  - Higher the value, Higher is the spread

Feature Engineering : Variable Clustering

Variables are clustered to understand if any two or more independent variables are correlated among each other

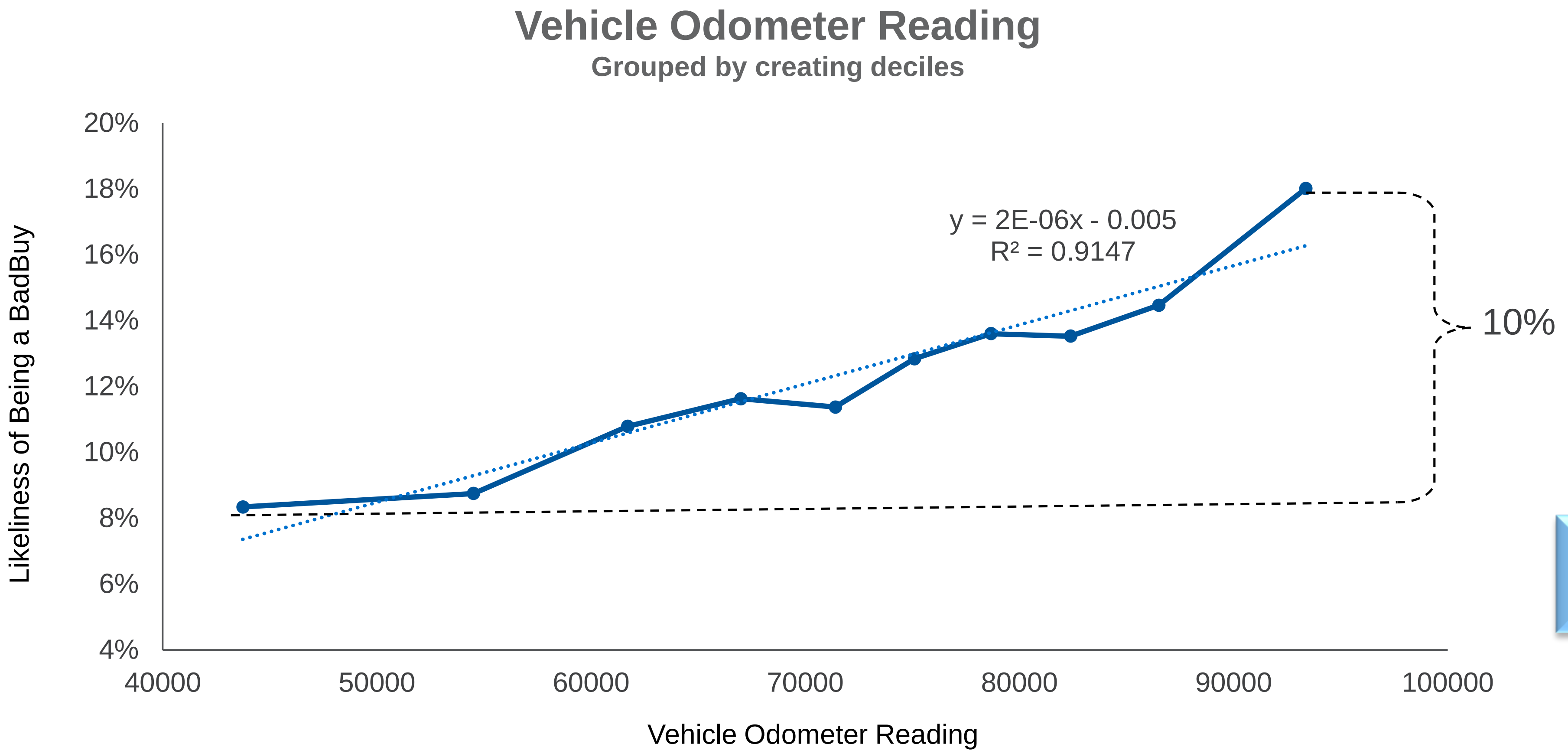
Multivariate analysis will be biased if any explanatory variable is correlated with other explanatory variable in the model (Multicollinearity Problem); This also assists in variable selection

- **Clustered the variables based on their correlation with other variables**
  - Variables in the same cluster are correlated among each other and have smaller distance between their mean
  - Variables in the different cluster are not correlated among each other and will have larger distance between their mean
  - Chose the solution with Max cluster in order to ensure that the model estimates won't be affected by multicollinearity
  - Other Advantages: It's an efficient statistical way to reduce the number of variables

- **Selected one or two variables per cluster and further visually studied them for their relationship with being a lemon purchase**

Feature Engineering: Relationship Assessment

Evaluated relationship of each element with the chance of being a lemon purchase (IsBadBuy), one at a time



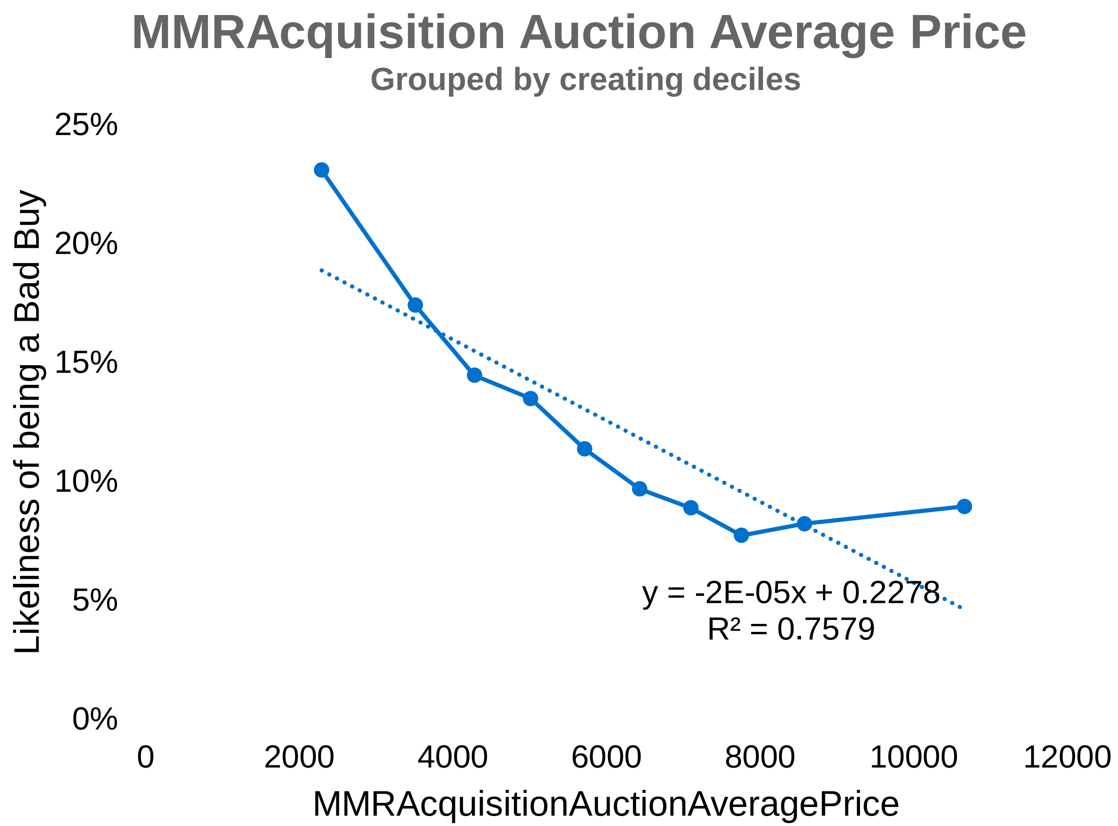
Higher the Separation, and smoother the relationship – Higher the explanatory power for the variable in the model

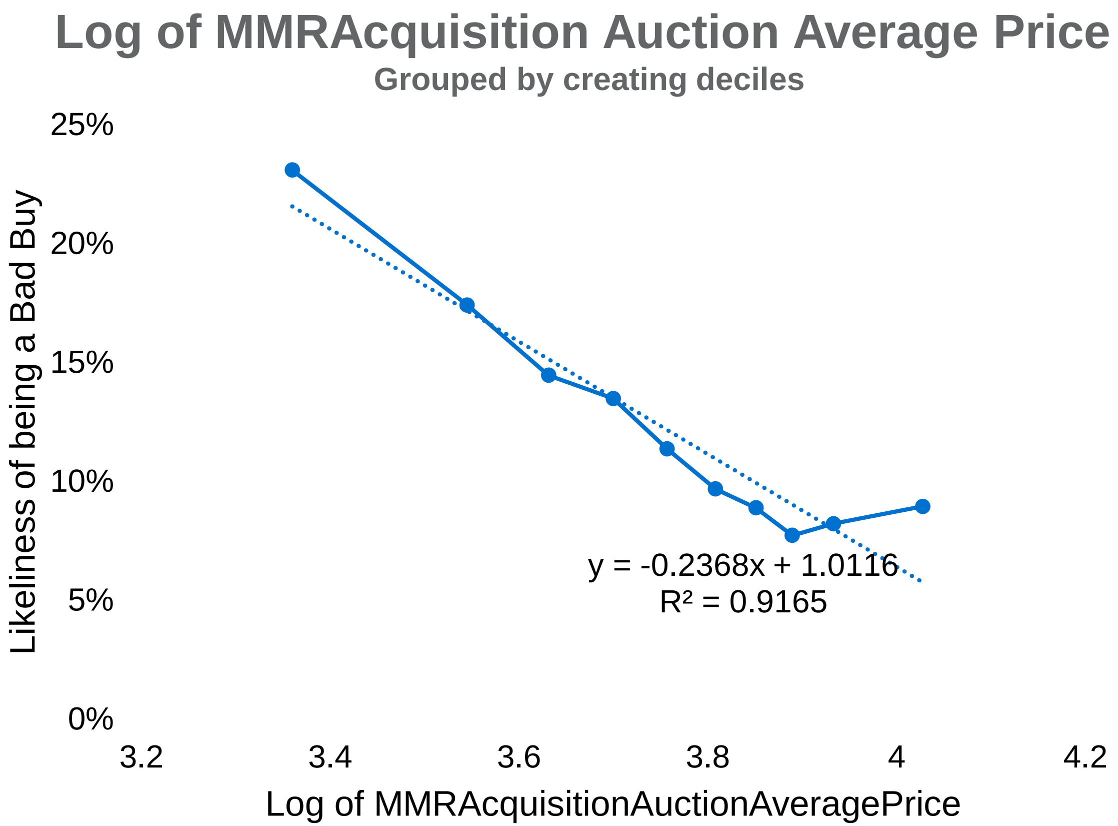
Similar analysis were conducted for all the other variables

Chance of a purchase being lemon increases by increase in mileage odometer reading

Feature Engineering: Variable Transformation

Variables were transformed to make the relationship with IsBadBuy simple and smoother



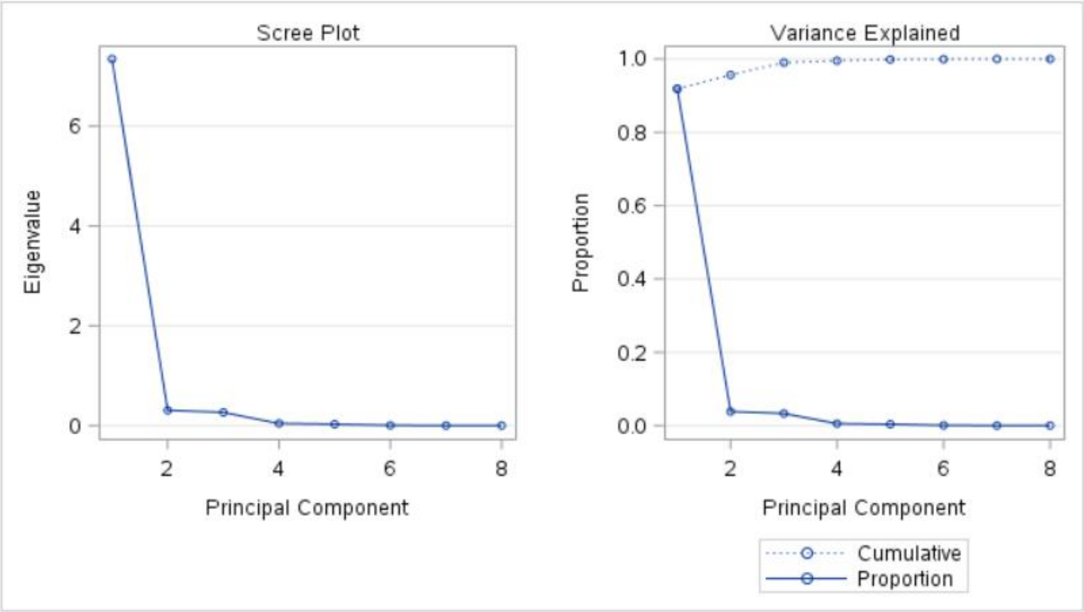


Log transformation of this variable smoothens the relationship with IsBadBuy, which ultimately results in a better fit of the model

Similar analysis were conducted for all the other variables

Feature Engineering : Principal Component Analysis

Auction price related variables were correlated among each other hence PCA was applied to create a smaller set of components and grab the majority of the variation within the data available



PCA is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components

- **The first principal component accounts for ~90% of the variance**
  - Tested the first component as potential input for the model

Roadmap to Achieve our Goal

The process below were used to evaluate the data; each and every variable was studied for their statistical value in both univariate and multivariate environment

Univariate Analysis

Data Processing | **Feature Engineering**

Ingest	Processing	Impute	Assess	Go/No Go
Read-in data, merged JSON with CSV, De-Duped	Translated, created new variables to assess usability	Impute missing values with appropriate information for better use of data	Measure statistical value and relationship with lemon purchases	Determine strategic and statistical value

Multi-variate Analysis (Model Building)

Ingest	Train	Assess	Validate	Score
Ingest the variables that passed round 1 and use learnings towards multivariate analysis	Train the model with the variables studied and refine it in a multivariate setting	Understand the outputs from the model and compare it against the univariate analysis performed	Validate the model on the validation sample and refine the model to reduce variance & bias in the prediction	Score the test dataset with the best trained model with the balanced variance & bias in prediction

Multivariate Modeling: Approach

Variables shortlisted will be ingested in the model to estimate the chance of a purchase being a bad buy

- **Dependent Variable: IsBadBuy**

- **Sample:**
  - Train dataset: 70% of Train_nw
  - Validation dataset: 30% of Train_nw

- **Modeling Approach: Logistic Regression**
  - Evaluate the statistical significance of the variables in the model
  - Evaluate the directional meaning of coefficients from the trend plotted during univariate analysis
  - Evaluate the statistically significant variables for multicollinearity

- **Evaluation Approach**
  - Score the valuation data with the coefficients from the model
  - Decile (equal groups of 10) the products by the descending order of predicted performance
  - Evaluate the separation & smoothness of the lift curve
  - Calculate RMSE (Root Mean Square Error) to measure the accuracy of the model

Multivariate Modeling: Logistic Regression Coefficients

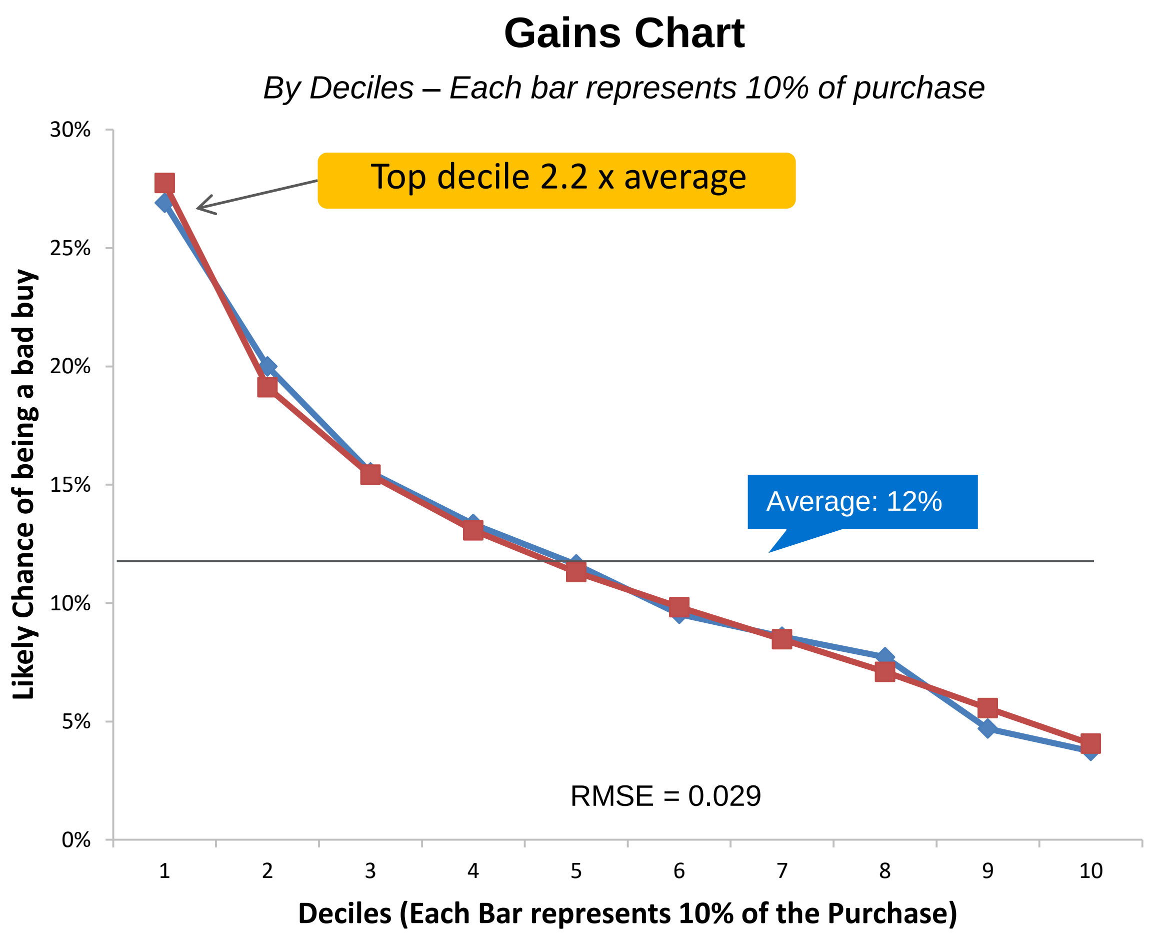
Variables in the model is ordered based on their importance

Analysis of Maximum Likelihood Estimates							
Parameter	**Label**	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept		1	5.4575	0.6197	78	0%	
WheelType_Covers	Wheel Type = 'Covers'	1	-0.6428	0.0379	288	0%	-0.1765
veh_age	Vehicle Age	1	0.1688	0.0109	238	0%	0.1588
log_vehbcost	Log of VehBCost	1	-0.9472	0.0666	202	0%	-0.1433
auction_ADESA	Auction Provider = 'ADESA'	1	0.3058	0.0385	63	0%	0.0673
tob3brand_GM	GM Brand Cars	1	-0.186	0.0357	27	0%	-0.0488
vehodo_10k	Vehicle Odometer (Unit of 10000)	1	0.0445	0.0123	13	0%	0.0358
size_COMPACT	Compact Cars	1	0.1346	0.0542	6	1%	0.0221

- **Interpretation:**
  - Veh Age – Higher the age of the vehicle, high likely the vehicle is a bad buy
    - Increase in age of the vehicle by 1 unit increase the logodds of being a bad buy by 0.1755
  - Veh Cost – Lower the cost of the vehicle, high likely the vehicle is a bad buy
    - Increase in 1 log unit of cost in vehicle decreases the logodds of the being a bad buy by -0.9415

Model – Validation

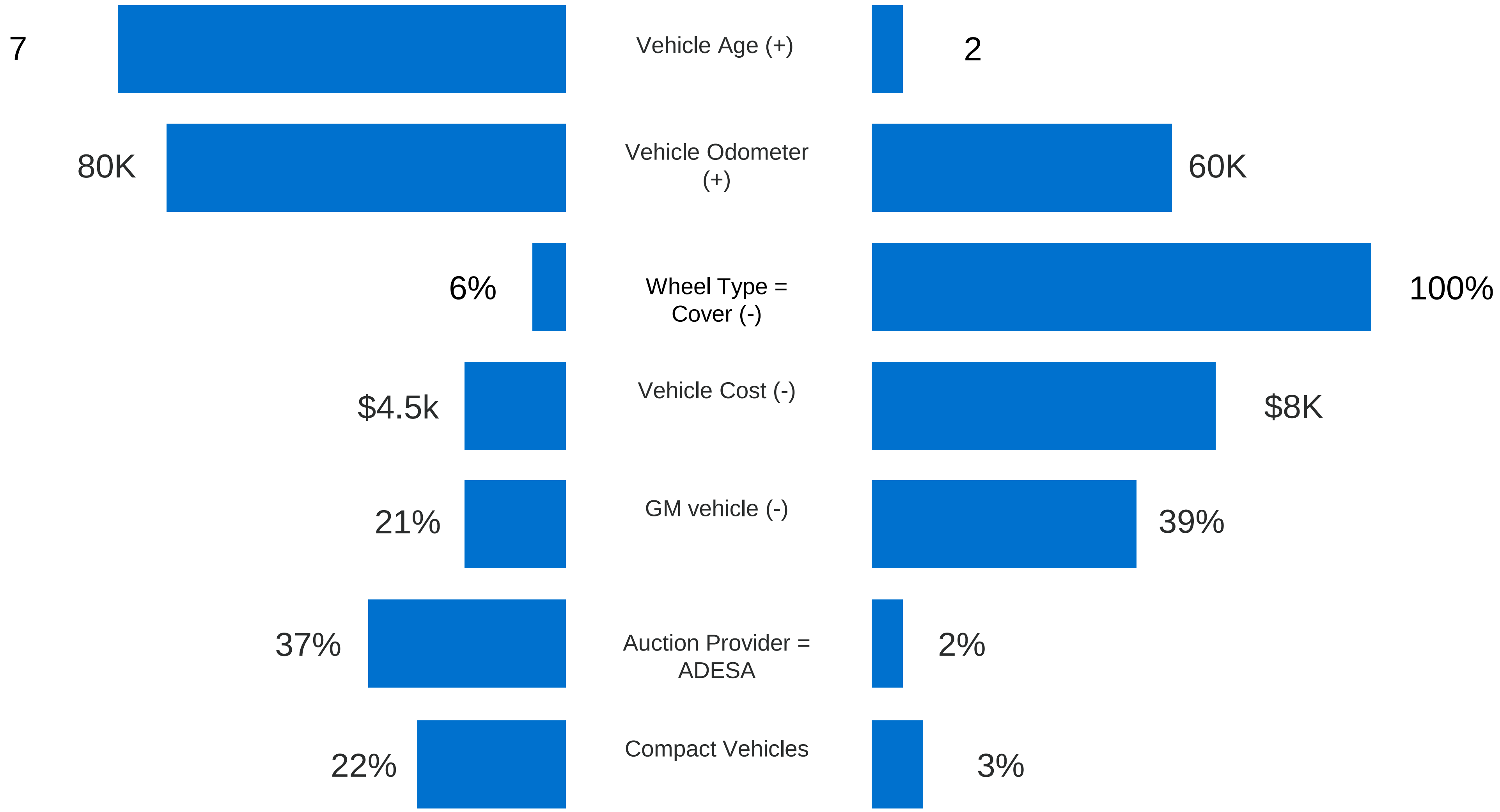
The model predicts the bad buy with low Root Mean Squared Error



Business Insight

High Likely Chance of being Bad Buy

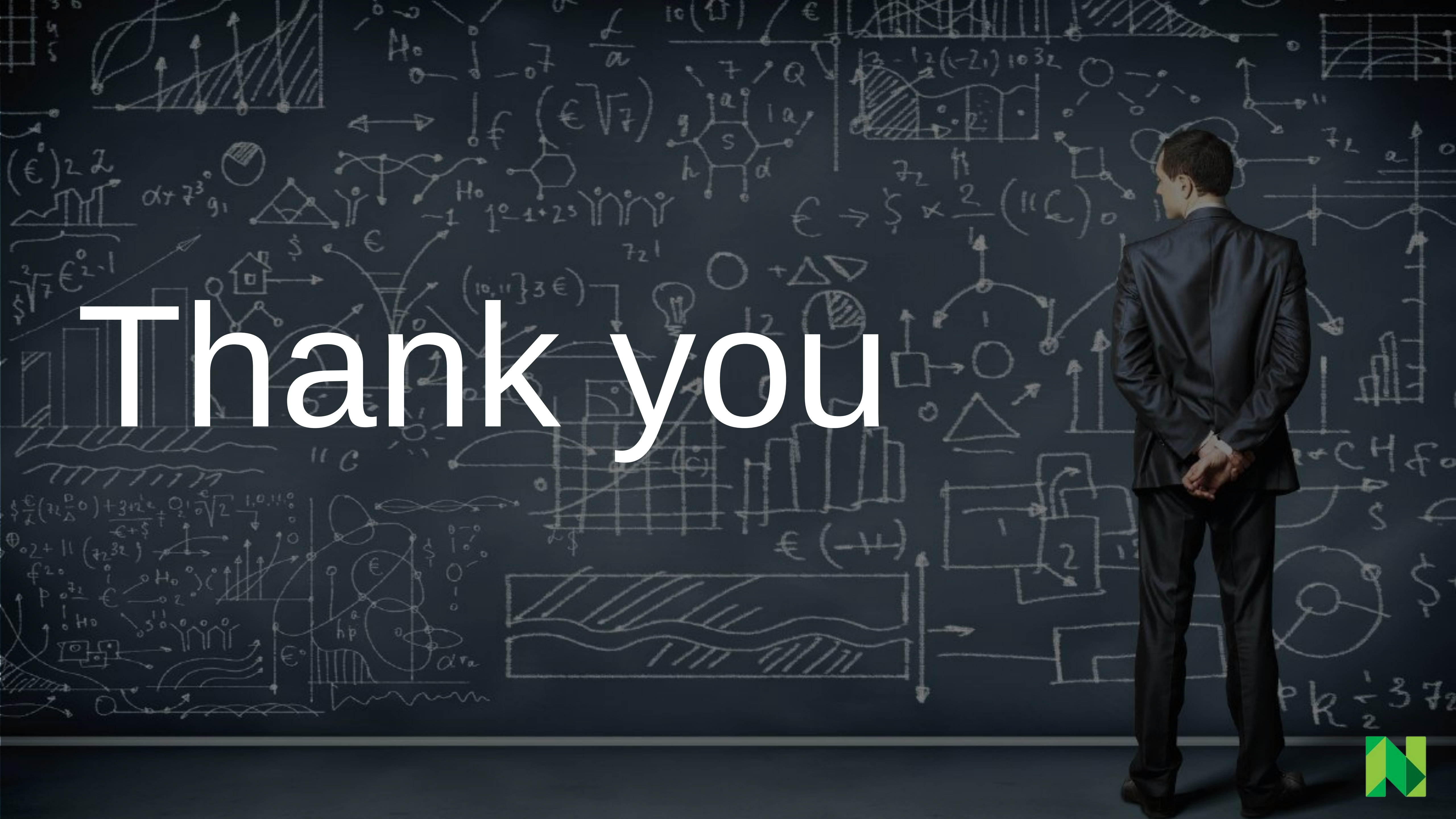
Low Likely Chance of being Bad Buy



Summary

- **In order to ensure that the vehicle purchased are not lemons, consider buying the vehicles with the below characteristics**
  - More recently purchased vehicle
  - Vehicle odometer no more than 60K
  - Vehicle with Cover wheel type
  - GM vehicles are more reliable
  - Compact vehicles are more reliable

- **Recommended Modeling Approach to test given additional time**
  - Ensemble methods (RandomForest & Gradient Boosting machine) – As both the model can capture randomness through sample and features, we could discover more complex relationship



Thank you