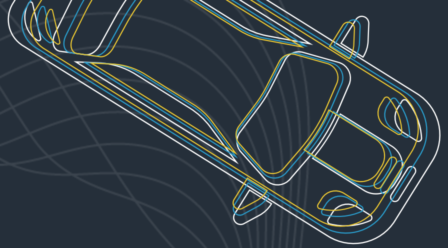
# Carvana - IsBadBuy?

## Group 8

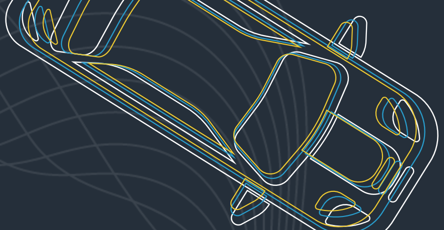
Emely Callejas, Ashley Cortez,
Rithvik V Sourab, Angelica Verduzco

# Case Background & Problem



- Established in 2012, Carvana Co. provides an online service for buying used cars, offering options for delivery or pickup in certain, mainly southern US, states.
- As of 2017, Carvana expanded to 60 metropolitan markets with nearly 10,000 pre-owned vehicles.
- Carvana’s growth strategy is dependent on integrated supply chain, an exclusive software system and data analytics
- The main focus on the case revolves around using a logistic regression for predictive analysis to determine primary factors influencing whether a car is a bad buy(IsBadBuy).

# Categorical Variables Available in the Data Set



- **Auction**: Auction provider at which the vehicle was purchased
- **VehYear**: Manufacturer’s year of the vehicle
- **Make**: Vehicle Manufacturer
- **Model**: Vehicle Model
- **Trim**: Vehicle Trim Level
- **SubModel**: Vehicle Submodel
- **Color**: Vehicle Color
- **Transmission**: Vehicle transmission type (Automatic, Manual)
- **WheelTypeID**: The type id of the vehicle wheel
- **WheelType**: The vehicle wheel type description (Alloy, Covers)
- **Nationality**: The Manufacturer’s Country
- **Size**: The size category of the vehicle (Compact, SUV, etc.)
- **TopThreeAmericanName**: Identifies if the manufacturer is one of the top three American manufacturers

● ● ●

- **PRIMEUNIT**: Identifies if the vehicle would have a higher demand than standard price
- **AUCGUART**: The level guarantee provided by auction for the vehicle (Green Light - Guaranteed/arbitrable, Yellow Light - caution/issue, red light - sold as is)
- **VNST**: State where the car was purchased
- **IsOnlineSale:** Identifies if the vehicle was originally purchased online
- **VNZIP1**: Zipcode where the car was purchased
- **PurchDate**: The date the vehicle was purchased at Auction

***Variables in Yellow = Used in Logistic Regression***

# Numerical Variables Available in the Data Set

- **IsBadBuy**: Identifies if the kicked vehicle was an avoidable purchase
- **VehicleAge**: Year elapsed since manufacturer's year
- **MMRAcquisitionAuctionAveragePrice**: price for this vehicle in average condition at time of purchase
- **MMRAcquisitionAuctionCleanPrice**: price for this vehicle in the above Average condition at time of purchase
- **MMRAcquisitionRetailAveragePrice**: price for this vehicle in the retail market in average condition at time of purchase
- **MMRAcquisitionRetailCleanPrice**: price for this vehicle in the retail market in above average condition at time of purchase
- **MMRCurrentAuctionAveragePrice**: price for this vehicle in average condition as of current day
- **MMRCurrentAuctionCleanPrice**: price for this vehicle in the above condition as of current day
- **MMRCurrentRetailAveragePrice**: price for this vehicle in the retail market in average condition as of current day
- **MMRCurrentRetailCleanPrice**: price for this vehicle in the retail market in above average condition as of current day
- **VehBCost**: acquisition cost paid for the vehicle at time of purchase
- **WarranyCost**: Warranty price (term=36month and millage=36k)
- **BYRNO**: Unique number assigned to buyer that purchased the vehicle

***Variables in Yellow = Used in Logistic Regression***

# Variables Omitted from Logistic Regression

## PRIMEUNIT

**95% NULL Values**

## AUCGUART

**95% NULL Values**

## Auction, Color, Nationality, Trim, Sub Model, Transmission

**These fields had attributes that had significant amounts of Bad Buys, however they had low significance levels when added to logistic regression model and were not influential when included in the Odds Ratio analysis**

# Variables Focused on for Logistic Regression

**Approach to Selecting Independent Variables:**

- Conduct exploratory research of training dataset by using Excel and Tableau visualizations to determine which attributes within each categorical variable contain the highest count of Bad Buys
- Select specific attributes within each variable to include in logistic regression
  - E.g., Make - Chevrolet, Model - PT Cruiser, States - TX, FL, etc.
- Each independent variable was analyzed to check its significance level based on p-values
- Odds Ratio analysis was also used to determine a variable's influence on the dependent variable.

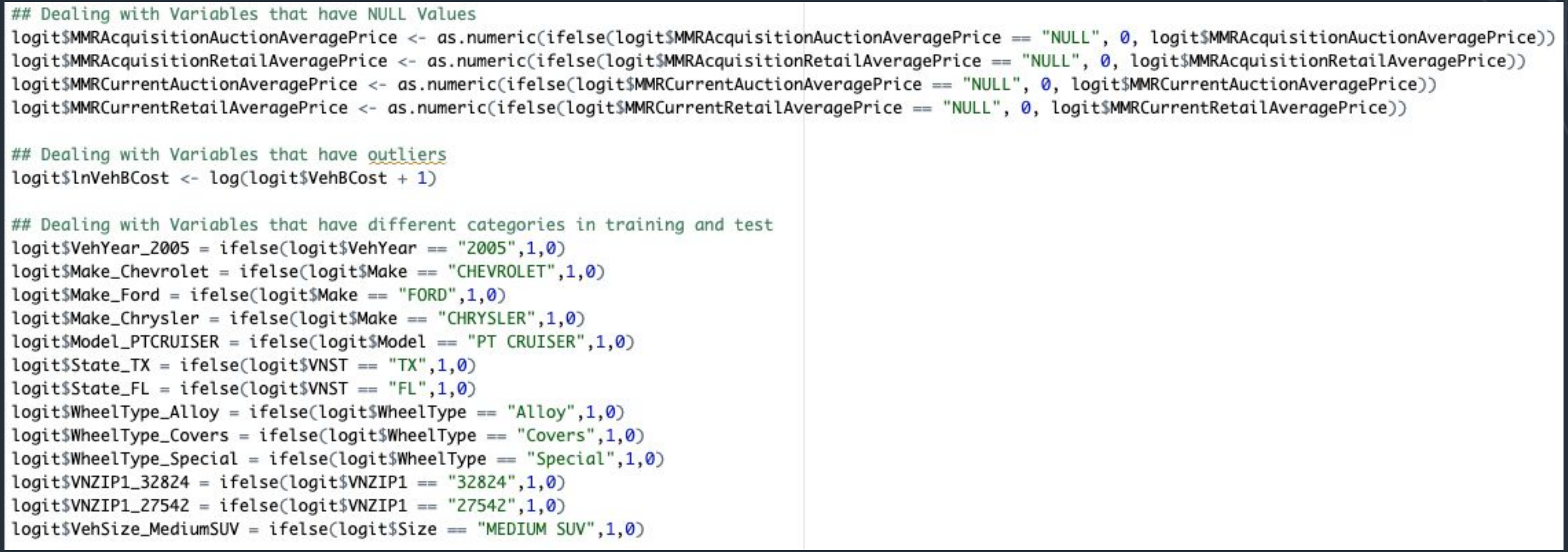
***See slide 7 for more details***

**Dependent Variable**
IsBadBuy

**Independent Variables**
VehicleAge
lnVehBCost
VehYear_2005
Make_Chevrolet
Model_PTCRUISER
State_TX
State_FL
WheelType_Alloy
WheelType_Covers
WheelType_Special
VNZIP1_27542
VehSize_MediumSUV
MMRCurrentRetailAveragePrice
MMRCurrentAuctionAveragePrice
MMRAcquisitionRetailAveragePrice
MMRAcquisitionAuctionAveragePrice

***Yellow = Specific attribute within variable***

# Transform & Clean Data



```
## Dealing with Variables that have NULL Values
logit$MMRAcquisitionAuctionAveragePrice <- as.numeric(ifelse(logit$MMRAcquisitionAuctionAveragePrice == "NULL", 0, logit$MMRAcquisitionAuctionAveragePrice))
logit$MMRAcquisitionRetailAveragePrice <- as.numeric(ifelse(logit$MMRAcquisitionRetailAveragePrice == "NULL", 0, logit$MMRAcquisitionRetailAveragePrice))
logit$MMRCurrentAuctionAveragePrice <- as.numeric(ifelse(logit$MMRCurrentAuctionAveragePrice == "NULL", 0, logit$MMRCurrentAuctionAveragePrice))
logit$MMRCurrentRetailAveragePrice <- as.numeric(ifelse(logit$MMRCurrentRetailAveragePrice == "NULL", 0, logit$MMRCurrentRetailAveragePrice))

## Dealing with Variables that have outliers
logit$lnVehBCost <- log(logit$VehBCost + 1)

## Dealing with Variables that have different categories in training and test
logit$VehYear_2005 = ifelse(logit$VehYear == "2005",1,0)
logit$Make_Chevrolet = ifelse(logit$Make == "CHEVROLET",1,0)
logit$Make_Ford = ifelse(logit$Make == "FORD",1,0)
logit$Make_Chrysler = ifelse(logit$Make == "CHRYSLER",1,0)
logit$Model_PTCRUISER = ifelse(logit$Model == "PT CRUISER",1,0)
logit$State_TX = ifelse(logit$VNST == "TX",1,0)
logit$State_FL = ifelse(logit$VNST == "FL",1,0)
logit$WheelType_Alloy = ifelse(logit$WheelType == "Alloy",1,0)
logit$WheelType_Covers = ifelse(logit$WheelType == "Covers",1,0)
logit$WheelType_Special = ifelse(logit$WheelType == "Special",1,0)
logit$VNZIP1_32824 = ifelse(logit$VNZIP1 == "32824",1,0)
logit$VNZIP1_27542 = ifelse(logit$VNZIP1 == "27542",1,0)
logit$VehSize_MediumSUV = ifelse(logit$Size == "MEDIUM SUV",1,0)
```

In the snapshot above, we are transforming and cleaning the variables that our included in our model. NULL variables in acquisition prices are replaced with 0. VehBCost is standardized to account for outliers. New columns are created for categorical variables to create new dummy variables.

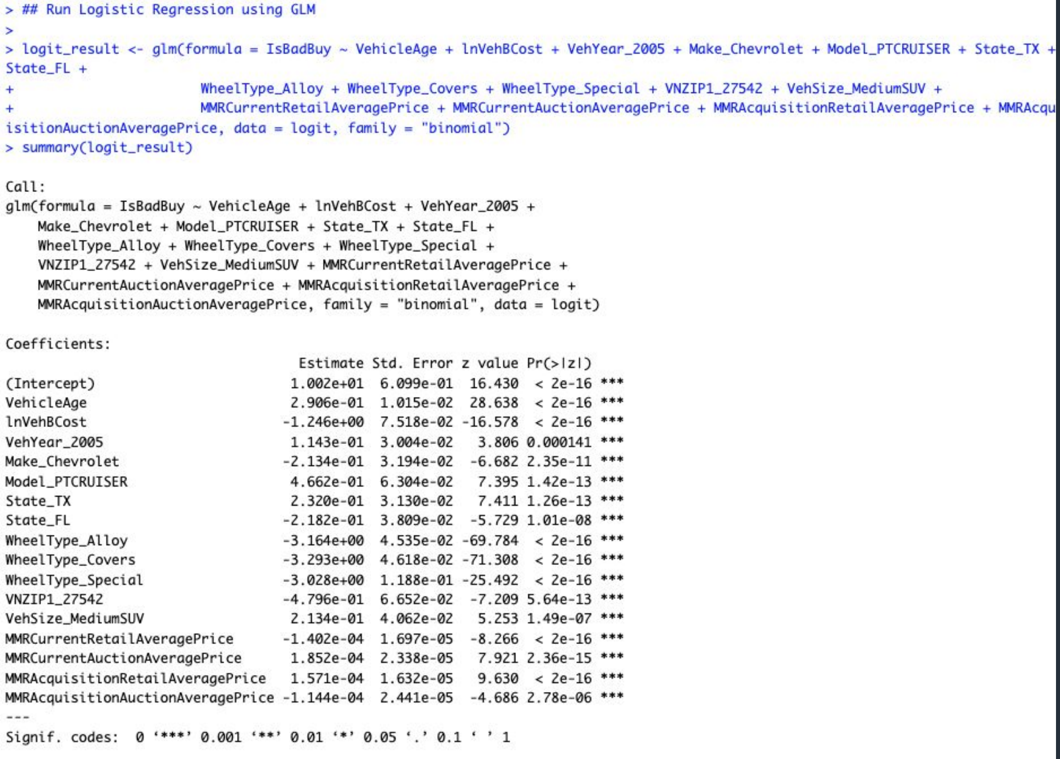
# Logistic Regression Model

When including categorical variables, we were sure to check that the p-value was 0.001 (indicated by ***)

With this analysis we were able to narrow down the most influential variables.

Variables that were not included (mentioned on slide 5) often had a p-value that was 0.05 or lower, indicating a low significance.

Wheel Type variables had the most surprising estimates and z-values.



```
> ## Run Logistic Regression using GLM
>
> logit_result <- glm(formula = IsBadBuy ~ VehicleAge + lnVehBCost + VehYear_2005 + Make_Chevrolet + Model_PTCRUISER + State_TX +
State_FL +
+                        WheelType_Alloy + WheelType_Covers + WheelType_Special + VNZIP1_27542 + VehSize_MediumSUV +
+                        MMRCurrentRetailAveragePrice + MMRCurrentAuctionAveragePrice + MMRAcquisitionRetailAveragePrice + MMRAcqu
isitionAuctionAveragePrice, data = logit, family = "binomial")
> summary(logit_result)

Call:
glm(formula = IsBadBuy ~ VehicleAge + lnVehBCost + VehYear_2005 +
    Make_Chevrolet + Model_PTCRUISER + State_TX + State_FL +
    WheelType_Alloy + WheelType_Covers + WheelType_Special +
    VNZIP1_27542 + VehSize_MediumSUV + MMRCurrentRetailAveragePrice +
    MMRCurrentAuctionAveragePrice + MMRAcquisitionRetailAveragePrice +
    MMRAcquisitionAuctionAveragePrice, family = "binomial", data = logit)

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.002e+01  6.099e-01  16.430  < 2e-16 ***
VehicleAge                        2.906e-01  1.015e-02  28.638  < 2e-16 ***
lnVehBCost                       -1.246e+00  7.518e-02 -16.578  < 2e-16 ***
VehYear_2005                      1.143e-01  3.004e-02   3.806 0.000141 ***
Make_Chevrolet                   -2.134e-01  3.194e-02  -6.682 2.35e-11 ***
Model_PTCRUISER                   4.662e-01  6.304e-02   7.395 1.42e-13 ***
State_TX                          2.320e-01  3.130e-02   7.411 1.26e-13 ***
State_FL                         -2.182e-01  3.809e-02  -5.729 1.01e-08 ***
WheelType_Alloy                  -3.164e+00  4.535e-02 -69.784  < 2e-16 ***
WheelType_Covers                 -3.293e+00  4.618e-02 -71.308  < 2e-16 ***
WheelType_Special                -3.028e+00  1.188e-01 -25.492  < 2e-16 ***
VNZIP1_27542                     -4.796e-01  6.652e-02  -7.209 5.64e-13 ***
VehSize_MediumSUV                 2.134e-01  4.062e-02   5.253 1.49e-07 ***
MMRCurrentRetailAveragePrice     -1.402e-04  1.697e-05  -8.266  < 2e-16 ***
MMRCurrentAuctionAveragePrice     1.852e-04  2.338e-05   7.921 2.36e-15 ***
MMRAcquisitionRetailAveragePrice  1.571e-04  1.632e-05   9.630  < 2e-16 ***
MMRAcquisitionAuctionAveragePrice -1.144e-04  2.441e-05  -4.686 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

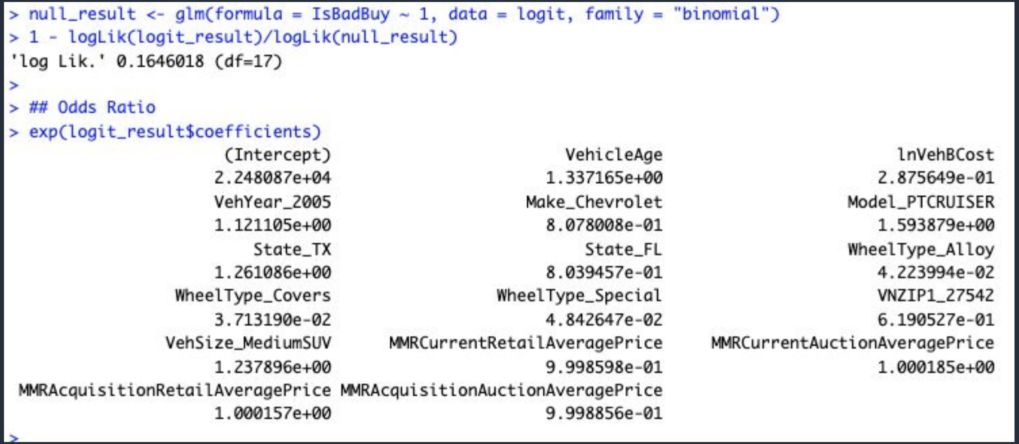
# Pseudo R-Square and Odds Ratio

We also checked the fit of the model with the Pseudo R-Square.

- Through trial and error of adding and removing variables in the model, we found that this is the highest r-square value from the various combinations

The odds ratio is the exponential of the coefficients and can be interpreted as such:

- For every unit increase of each variable, the odds of the dependent variable changing are influenced (positively or negatively)
- E.g., for every increase in VehicleAge, the odds of Bad Buy increases by 33.7%



```
> null_result <- glm(formula = IsBadBuy ~ 1, data = logit, family = "binomial")
> 1 - logLik(logit_result)/logLik(null_result)
'log Lik.' 0.1646018 (df=17)
>
> ## Odds Ratio
> exp(logit_result$coefficients)
                       (Intercept)                         VehicleAge                          lnVehBCost 
                      2.248087e+04                       1.337165e+00                        2.875649e-01 
                      VehYear_2005                     Make_Chevrolet                    Model_PTCRUISER 
                      1.121105e+00                       8.078008e-01                        1.593879e+00 
                          State_TX                           State_FL                     WheelType_Alloy 
                      1.261086e+00                       8.039457e-01                        4.223994e-02 
                  WheelType_Covers                  WheelType_Special                        VNZIP1_27542 
                      3.713190e-02                       4.842647e-02                        6.190527e-01 
                  VehSize_MediumSUV          MMRCurrentRetailAveragePrice    MMRCurrentAuctionAveragePrice 
                      1.237896e+00                       9.998598e-01                        1.000185e+00 
 MMRAcquisitionRetailAveragePrice  MMRAcquisitionAuctionAveragePrice 
                      1.000157e+00                       9.998856e-01 
>
```

# Hit Rate Table

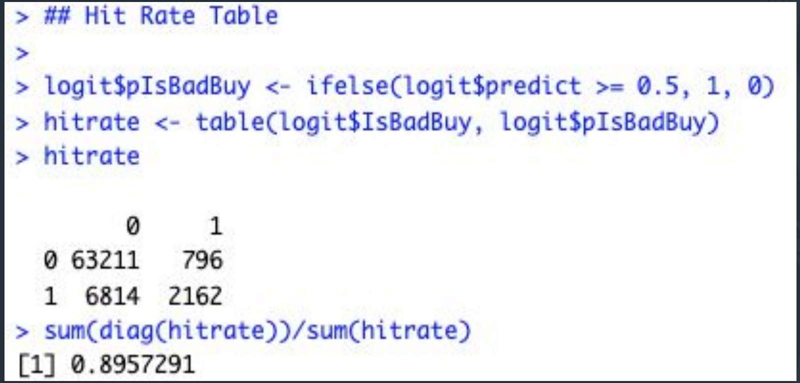
The Hit Rate Table offered us a clear view of what the model is predicting accurately and inaccurately.

**Insights**

- True Negatives: 63,211 - correctly predicted non-Bad Buys
- True Positives: 2,162 - correctly predicted Bad Buys
- False Positives: 796 - Non-Bad Buys incorrectly labeled as Bad Buys
- False Negatives: 6,814 - Bad Buys missed by the model

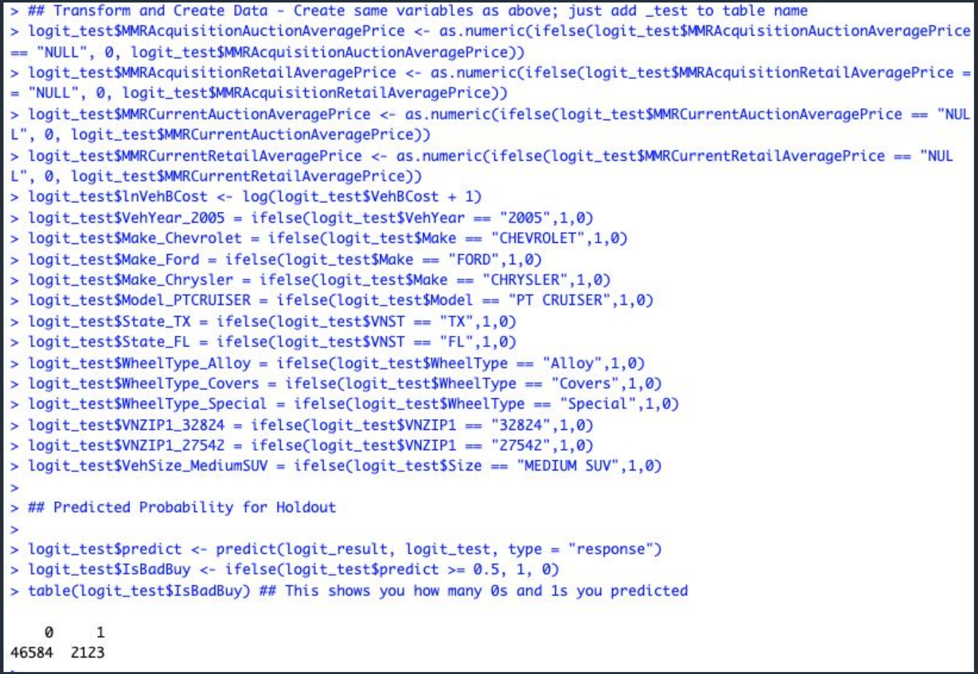
**Model Accuracy:** **89.57%**

- Indicates a high level of correct overall predictions



```
> ## Hit Rate Table
>
> logit$pIsBadBuy <- ifelse(logit$predict >= 0.5, 1, 0)
> hitrate <- table(logit$IsBadBuy, logit$pIsBadBuy)
> hitrate

        0     1
  0 63211   796
  1  6814  2162
> sum(diag(hitrate))/sum(hitrate)
[1] 0.8957291
```

# Predict Bad Buy on Test Data Set



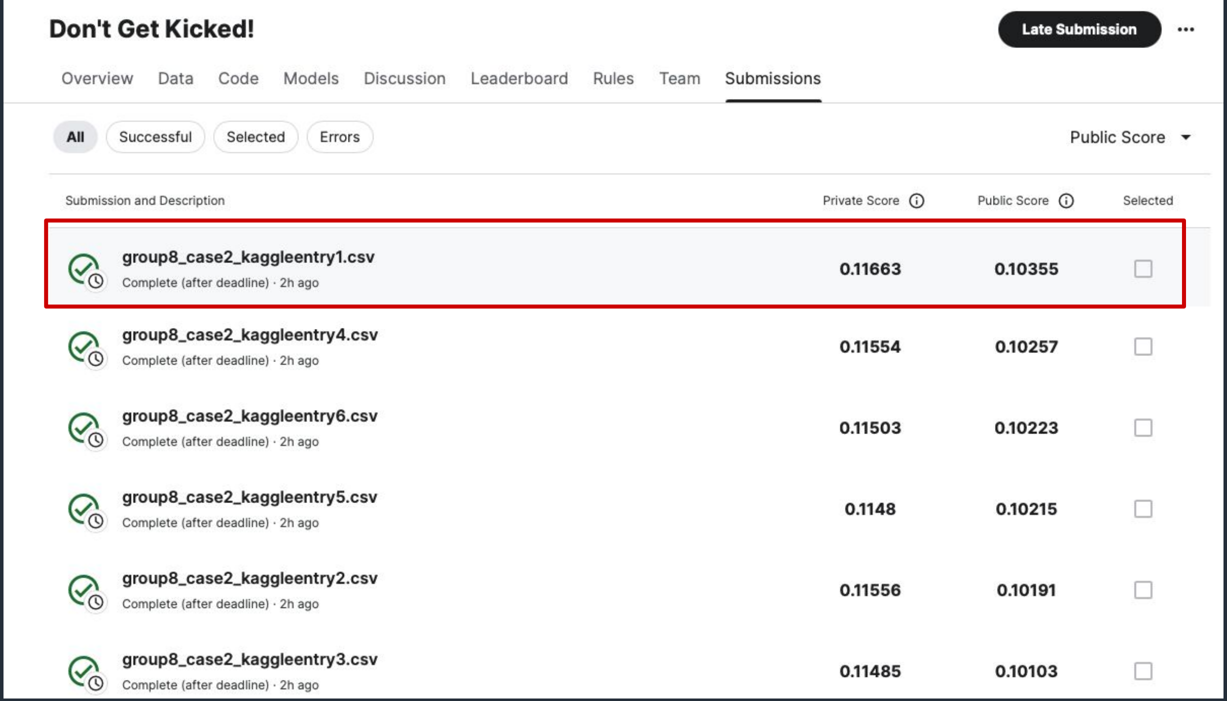
```
> ## Transform and Create Data - Create same variables as above; just add _test to table name
> logit_test$MMRAcquisitionAuctionAveragePrice <- as.numeric(ifelse(logit_test$MMRAcquisitionAuctionAveragePrice
== "NULL", 0, logit_test$MMRAcquisitionAuctionAveragePrice))
> logit_test$MMRAcquisitionRetailAveragePrice <- as.numeric(ifelse(logit_test$MMRAcquisitionRetailAveragePrice =
= "NULL", 0, logit_test$MMRAcquisitionRetailAveragePrice))
> logit_test$MMRCurrentAuctionAveragePrice <- as.numeric(ifelse(logit_test$MMRCurrentAuctionAveragePrice == "NUL
L", 0, logit_test$MMRCurrentAuctionAveragePrice))
> logit_test$MMRCurrentRetailAveragePrice <- as.numeric(ifelse(logit_test$MMRCurrentRetailAveragePrice == "NUL
L", 0, logit_test$MMRCurrentRetailAveragePrice))
> logit_test$lnVehBCost <- log(logit_test$VehBCost + 1)
> logit_test$VehYear_2005 = ifelse(logit_test$VehYear == "2005",1,0)
> logit_test$Make_Chevrolet = ifelse(logit_test$Make == "CHEVROLET",1,0)
> logit_test$Make_Ford = ifelse(logit_test$Make == "FORD",1,0)
> logit_test$Make_Chrysler = ifelse(logit_test$Make == "CHRYSLER",1,0)
> logit_test$Model_PTCRUISER = ifelse(logit_test$Model == "PT CRUISER",1,0)
> logit_test$State_TX = ifelse(logit_test$VNST == "TX",1,0)
> logit_test$State_FL = ifelse(logit_test$VNST == "FL",1,0)
> logit_test$WheelType_Alloy = ifelse(logit_test$WheelType == "Alloy",1,0)
> logit_test$WheelType_Covers = ifelse(logit_test$WheelType == "Covers",1,0)
> logit_test$WheelType_Special = ifelse(logit_test$WheelType == "Special",1,0)
> logit_test$VNZIP1_32824 = ifelse(logit_test$VNZIP1 == "32824",1,0)
> logit_test$VNZIP1_27542 = ifelse(logit_test$VNZIP1 == "27542",1,0)
> logit_test$VehSize_MediumSUV = ifelse(logit_test$Size == "MEDIUM SUV",1,0)
>
> ## Predicted Probability for Holdout
>
> logit_test$predict <- predict(logit_result, logit_test, type = "response")
> logit_test$IsBadBuy <- ifelse(logit_test$predict >= 0.5, 1, 0)
> table(logit_test$IsBadBuy) ## This shows you how many 0s and 1s you predicted

    0     1
46584  2123
```

With what the model learned from the training data, 2,123 Bad Buys were predicted within the test data.

This accounts for approximately **4.4%** of the 48,707 records in the test data set

In the training data, ~12% of the records were identified as Bad Buys. It is important to recall that the hit table on the previous slide shows an 89% accuracy rate, which may explain why there is a lower percentage of Bad Buys predicted in the test data set.

# Kaggle Submission Results



Don't Get Kicked!

| Submission and Description | Private Score | Public Score | Selected |
|---|---|---|---|
| group8_case2_kaggleentry1.csv<br>Complete (after deadline) · 2h ago | 0.11663 | 0.10355 | □ |
| group8_case2_kaggleentry4.csv<br>Complete (after deadline) · 2h ago | 0.11554 | 0.10257 | □ |
| group8_case2_kaggleentry6.csv<br>Complete (after deadline) · 2h ago | 0.11503 | 0.10223 | □ |
| group8_case2_kaggleentry5.csv<br>Complete (after deadline) · 2h ago | 0.1148 | 0.10215 | □ |
| group8_case2_kaggleentry2.csv<br>Complete (after deadline) · 2h ago | 0.11556 | 0.10191 | □ |
| group8_case2_kaggleentry3.csv<br>Complete (after deadline) · 2h ago | 0.11485 | 0.10103 | □ |

# Kaggle Submission Results - Conclusion

- We tried multiple combinations of independent variables to see how they'd affect our overall score in the Kaggle submission.
- The independent variables we included in this presentation resulted in the highest score as well as the highest r-square
- Variables like VehicleAge, Make, Model, and State are included because they are known to affect a car's value and reliability, which could indicate a higher chance of being a bad buy. Also, choosing states such as TX and FL since they are populous states and represent a big amount of the data. These all were expected variables.
- However, we have excluded variables like color or features with less impact with other features.
- Some of the surprising variables were as follows:
  - VehYear_2005: We believe that there was perhaps a recall on a certain car that made this specific year impactful
  - Wheel Type (Alloy, Covers, Special): We believe buyers think certain wheel types are associated with vehicle quality/care.