



# Machine Learning for People

Gus Lipkin  
2023-08-25



## Agenda

- Motivations
- Setting the Stage
- Preview the Data
- Create Some Models
- Compare Results and Efficacy
- Thinking Beyond
- Summing Up

# Motivations

## How can I create a model that is easy to use and understand?

- You're building a model for a less technical team, they need to know what's happening
- You just care about the measured impact of the independent variables and not as much about the dependent
- Oftentimes these are business questions rather than technology questions

# Setting the Stage

## My Old Car



How many miles do I have remaining?

## The Instrument Cluster



October 1st, 2020, Costco Orlando, Very worried about running out

20 city

25 highway

22 combined



**My Google Form**

### Gas Tank

Distance traveled \*  
Your answer \_\_\_\_\_


How many gallons? \*  
Your answer \_\_\_\_\_

Price per gallon \*  
Your answer \_\_\_\_\_

Is the light on? \*  
 True  
 False

Fuel tank gauge level \*  
Choose ▾

Choose  
-.25  
0  
.25  
.5  
.75  
1  
1.25  
1.5  
1.75  
2  
2.25  
2.5  
2.75  
3  
3.25  
3.5  
3.75  
4



173 rows of data

Be able to compute my miles remaining while driving

## Taking a Look at the Data

| Timestamp           | Gallons | PricePerGallon | Light | GaugeLevel | Distance |
|---------------------|---------|----------------|-------|------------|----------|
| 2020-11-25 11:54:15 | 10.000  | \$1.779        | FALSE | 1.25       | 256.8    |
| 2020-11-25 11:54:51 | 8.986   | \$1.999        | FALSE | 1.5        | 255.1    |
| 2020-11-25 11:55:30 | 8.530   | \$1.839        | FALSE | 1.75       | 233.4    |
| 2020-11-25 11:56:09 | 10.126  | \$1.679        | FALSE | 1.25       | 253.8    |
| 2020-11-25 11:57:15 | 6.085   | \$1.739        | FALSE | 2.75       | 143.7    |
| 2020-12-12 06:43:23 | 12.689  | \$1.779        | TRUE  | 0.25       | 231.2    |

## Computed Columns

| TotalCost | GallonsRemaining | MPG    | DollarsPerMile | ActualTankLevel | PercentError | MilesRemaining |
|-----------|------------------|--------|----------------|-----------------|--------------|----------------|
| \$17.790  | 5.900            | 25.680 | \$0.069        | 1.484           | 15.784%      | 151.512        |
| \$17.963  | 6.914            | 28.389 | \$0.070        | 1.739           | 13.762%      | 196.279        |
| \$15.687  | 7.370            | 27.362 | \$0.067        | 1.854           | 5.614%       | 201.660        |
| \$17.002  | 5.774            | 25.064 | \$0.067        | 1.453           | 13.946%      | 144.721        |
| \$10.582  | 9.815            | 23.615 | \$0.074        | 2.469           | 11.373%      | 231.786        |
| \$22.574  | 3.211            | 18.221 | \$0.098        | 0.808           | 69.052%      | 58.506         |

15.9 gallons

Rescale gallons remaining to get actual tank

Percent error in tank level

Miles remaining = gallons remaining \* MPG

# Thinking About Machine Learning

## Goals

- Don't run out of gas
- Be able to compute my miles remaining while driving
- Have an accurate estimate of miles remaining at any moment
- Share my findings in an easy to understand way

Be able to compute my miles remaining while driving

## Models

Likely to work:

- Produces continuous numeric output
- Error can be calculated

Needs modification:

- Produces discrete numeric output
- Produces probabilistic output
- Error is harder to calculate

## Model Metrics

- Estimated Miles Remaining
  - RMSE: Root Mean Square Error
    - How much of a buffer do I want?
- Percent Accuracy

Likely To Work



## Neural Net

MilesRemaining  $\sim$  GaugeLevel + Light + Distance

Uses:

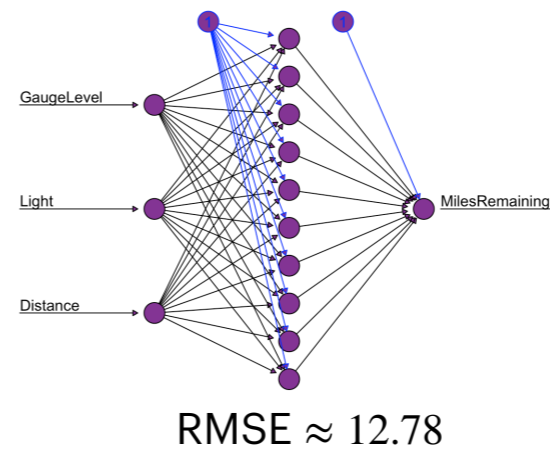
- All available variables
- 1000 epochs
- 10 hidden units

Pros:

- Very accurate
- Very cool to use

Cons:

- Not easy to do on the fly
- Difficult to communicate
- Computationally expensive



RMSE = root mean square error of miles remaining

## Simple Decision Tree

MilesRemaining  $\sim$  GaugeLevel + Light + Distance

Uses:

- All available variables
  - The model only actually uses GaugeLevel

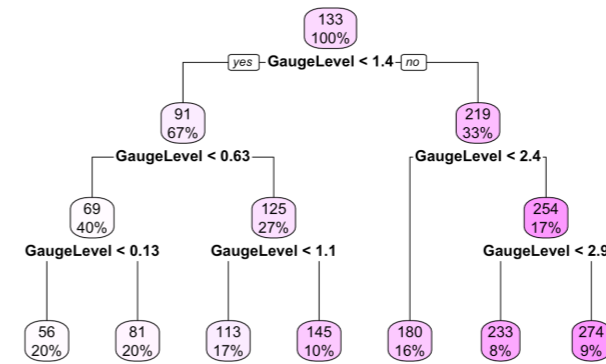
- 1 tree

Pros:

- Easy to understand
- Easy to execute

Cons:

- Could be better
- Distance + GaugeLevel implies MPG
- A little bit of memorization



RMSE  $\approx$  15.18

# Random Forest

MilesRemaining ~ GaugeLevel + Light + Distance

Uses:

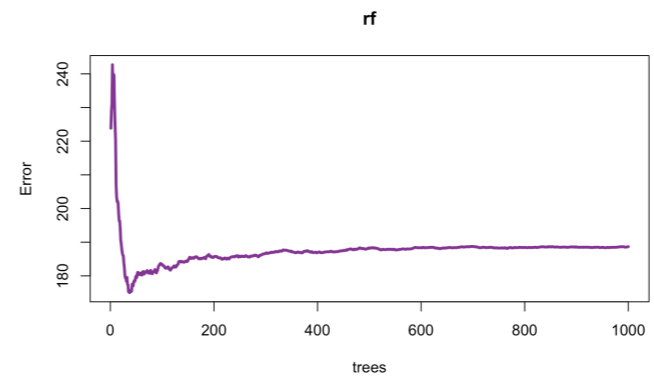
- All available variables
- 1000 trees

Pros:

- Very accurate
- Not computationally expensive

Cons:

- Not easy to do on the fly
- Difficult to communicate



RMSE  $\approx 7.3$

## Simple Linear Regression

MilesRemaining ~ GaugeLevel

Uses:

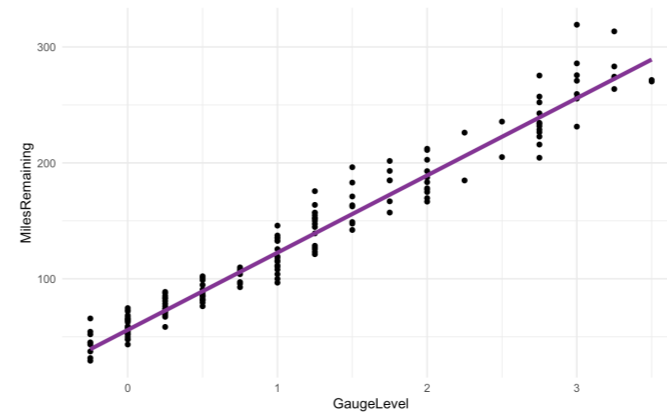
- GaugeLevel

Pros:

- Easy to understand
- Easy to execute

Cons:

- Could be better
- Distance + GaugeLevel implies MPG
- Requires some mental math



$$MilesRemaining = 66.23 * GaugeLevel + 55.94$$

$$RMSE \approx 14.31$$

## Multiple Linear Regression

MilesRemaining ~ GaugeLevel + Light + Distance

Uses:

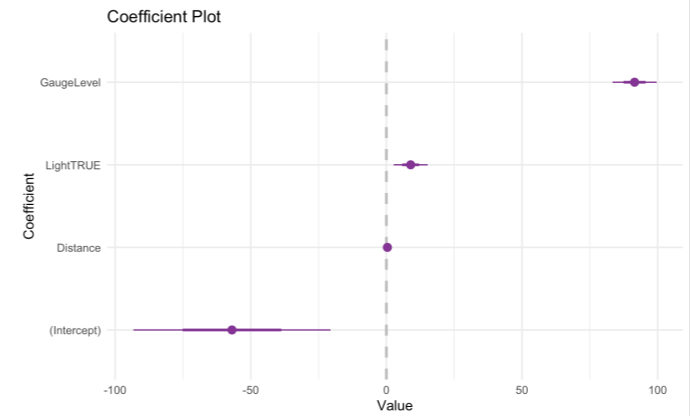
- All available variables

Pros:

- Easy to communicate
- Very Accurate

Cons:

- Lots of mental math
- Not as easy to understand



MilesRemaining = 91.496 \* GaugeLevel + 8.949 \* Light + 0.367 \* Distance - 56.936

RMSE  $\approx$  12.85

## Needs Modification

## K-Means

MilesRemaining ~ GaugeLevel + Light + Distance

Uses:

- All variables
- 10 centers

Pros:

- Easy to execute

Cons:

- Lots of human factors in interpretation
- Difficult to choose a row
- Difficult to communicate
- No good measure of accuracy

Centroids

| MilesRemaining | GaugeLevel | Light | Distance |
|----------------|------------|-------|----------|
| 47.267         | -0.141     | 1.000 | 301.225  |
| 65.871         | 0.083      | 0.963 | 285.096  |
| 86.153         | 0.500      | 0.208 | 264.767  |
| 90.856         | 0.438      | 0.250 | 289.500  |
| 116.424        | 0.908      | 0.000 | 247.147  |
| 128.073        | 1.232      | 0.000 | 213.479  |
| 167.513        | 1.395      | 0.000 | 246.174  |
| 181.677        | 1.929      | 0.000 | 173.421  |
| 227.054        | 2.672      | 0.000 | 128.969  |
| 276.483        | 3.133      | 0.000 | 98.513   |

# Logistic Regression

MilesRemaining ~ GaugeLevel

Uses:

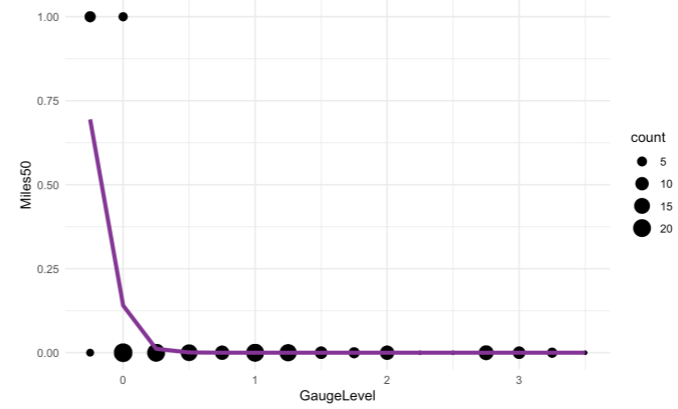
- GaugeLevel
- MilesRemaining converted to  $\leq 50$  miles remaining

Pros:

- Easy to understand
- Relatively accurate

Cons:

- Almost the same as the GaugeLevel
- You either run out of gas or you don't...
  - MilesRemaining is  $\leq 50$  at  $-.25$  GaugeLevel

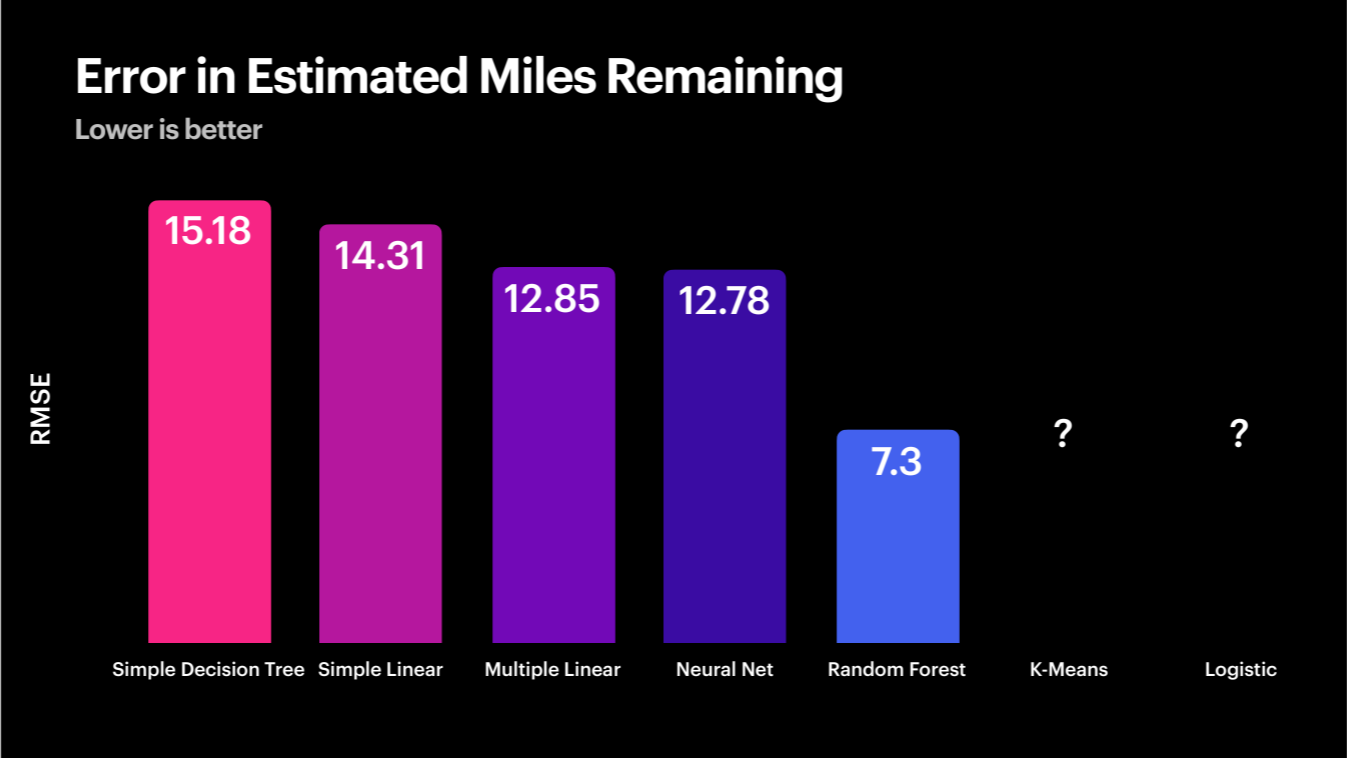


$$\text{Miles50} = \frac{.095}{\text{GaugeLevel} + .267} - 1$$

Accuracy  $\approx 95.93\%$



# Model Comparisons



Missing k-means and logistic

## Ease of Use

Higher is better



Simple Decision Tree

Simple Linear

Multiple Linear

Neural Net

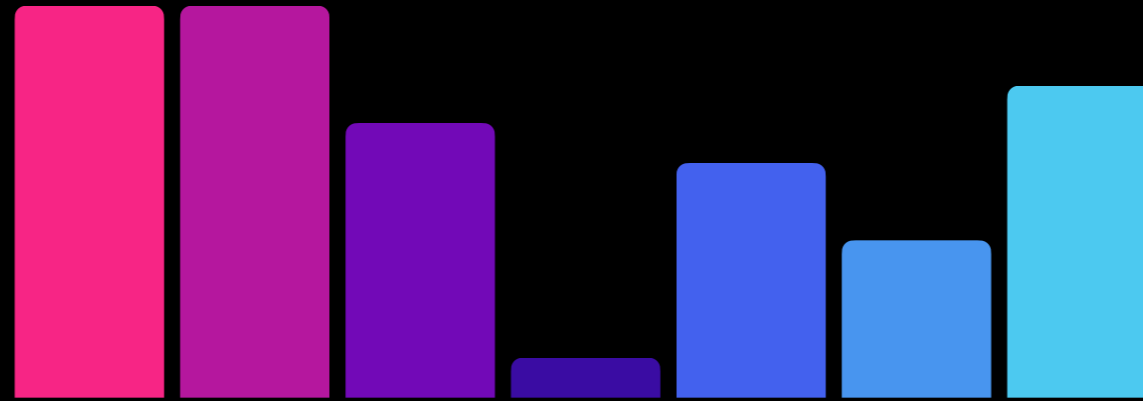
Random Forest

K-Means

Logistic

## Ease of Communication

Higher is better



Simple Decision Tree

Simple Linear

Multiple Linear

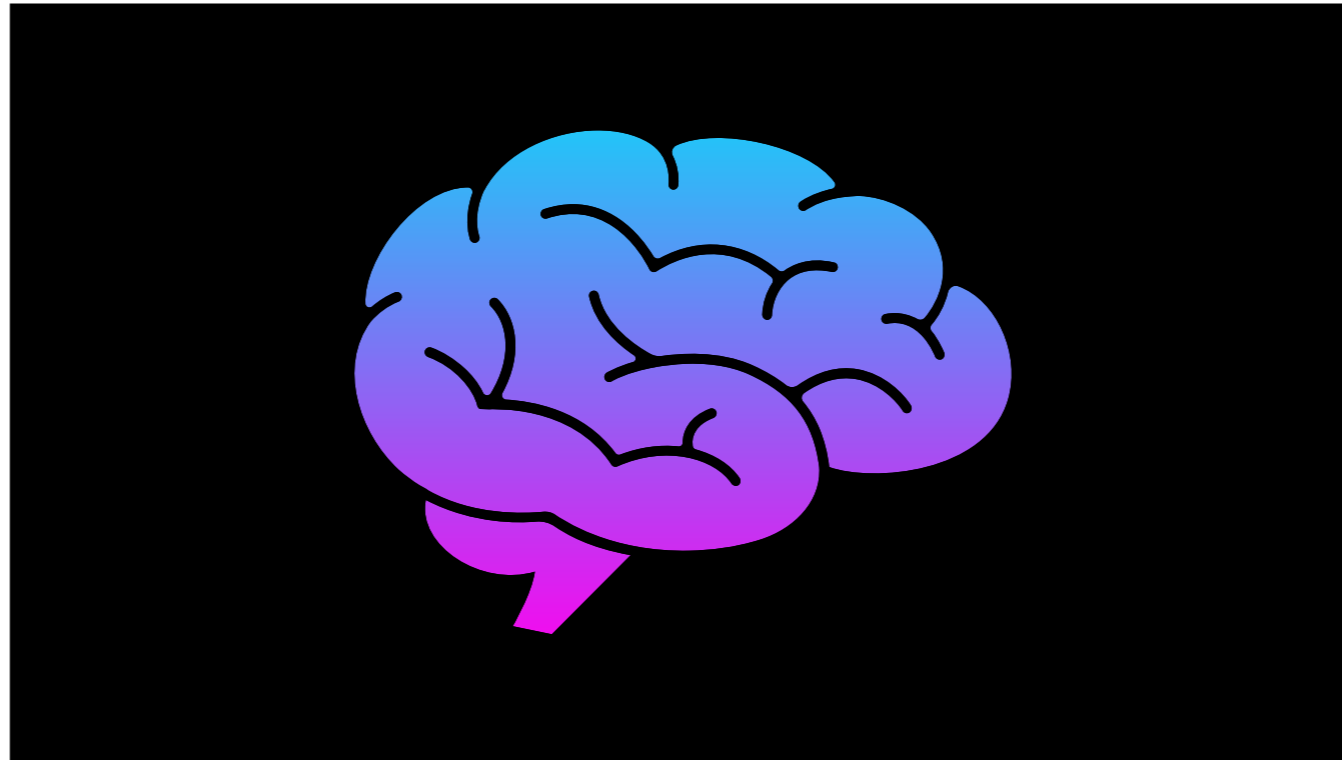
Neural Net

Random Forest

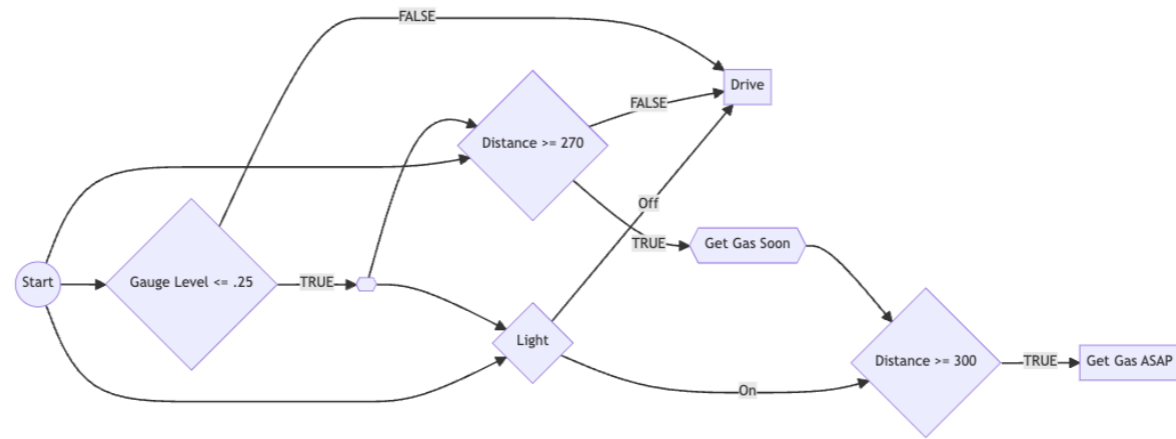
K-Means

Logistic

What did I use to estimate my miles remaining?

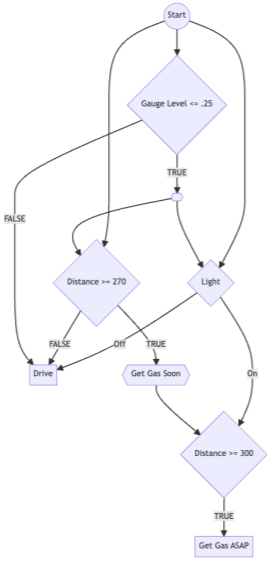


## It's Basically Decision Tree



# Meets K-Means Table

Centroids



| MilesRemaining | GaugeLevel | Light | Distance |
|----------------|------------|-------|----------|
| 47.267         | -0.141     | 1.000 | 301.225  |
| 65.871         | 0.083      | 0.963 | 285.096  |
| 86.153         | 0.500      | 0.208 | 264.767  |
| 90.856         | 0.438      | 0.250 | 289.500  |
| 116.424        | 0.908      | 0.000 | 247.147  |
| 128.073        | 1.232      | 0.000 | 213.479  |
| 167.513        | 1.395      | 0.000 | 246.174  |
| 181.677        | 1.929      | 0.000 | 173.421  |
| 227.054        | 2.672      | 0.000 | 128.969  |
| 276.483        | 3.133      | 0.000 | 98.513   |



# Thinking Beyond the Gas Tank

## Hospital Patient Readmissions

What influences readmissions?

## Wait Times at Disney World

What influences wait times?

## Standardized Test Scores

What influences test scores?

## Summing Up

- Accuracy is good
  - But so is explainability and executability
- Sometimes it's okay to sacrifice some accuracy

As seen in Jared's talk, more complex models can be wasteful for marginal gains, and sometimes worse



# Machine Learning for People

Gus Lipkin  
2023-08-25