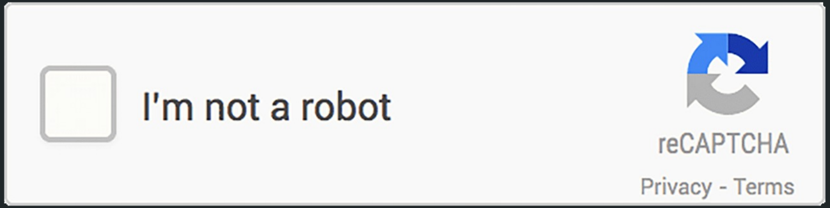
I’m not a robot vs a data scientist?

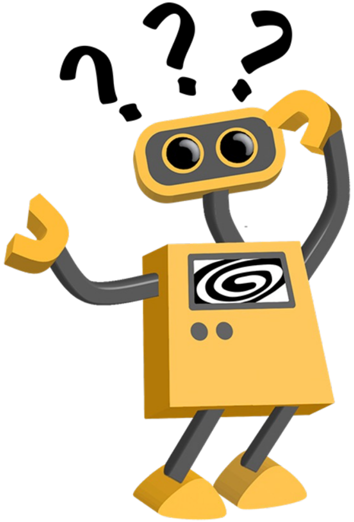
Utilizing machine learning principles and image classification methods to overcome website prompts asking for proof that …

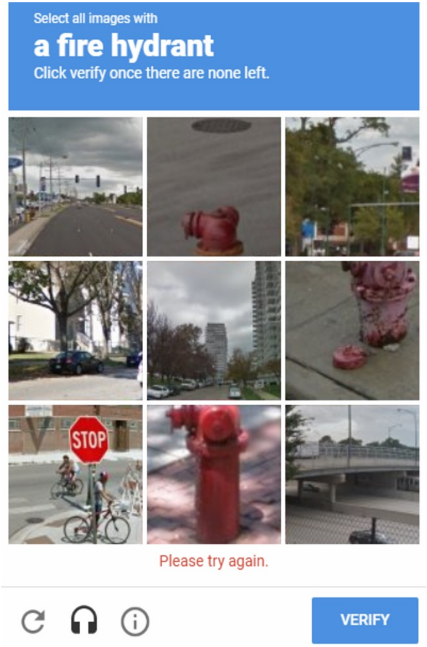


□ I'm not a robot

How effective are website protection measures

Can the computer determine whether a user is a human or not by assessing whether it can effectively identify if an image is an object or living thing using the CIFAR-10 data set?





Select all images with

a fire hydrant

Click verify once there are none left.

Please try again.

Data Collection, EDA, & Data Pre-processing

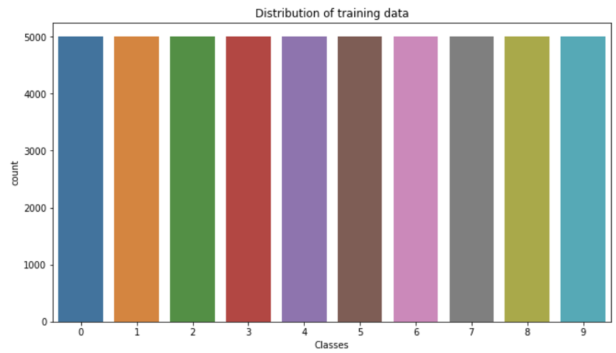
Data Collection

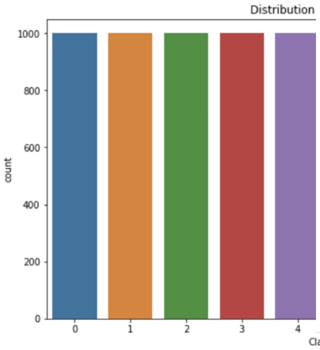
CIFAR-10 image data set, a widely used dataset for machine learning and image recognition.

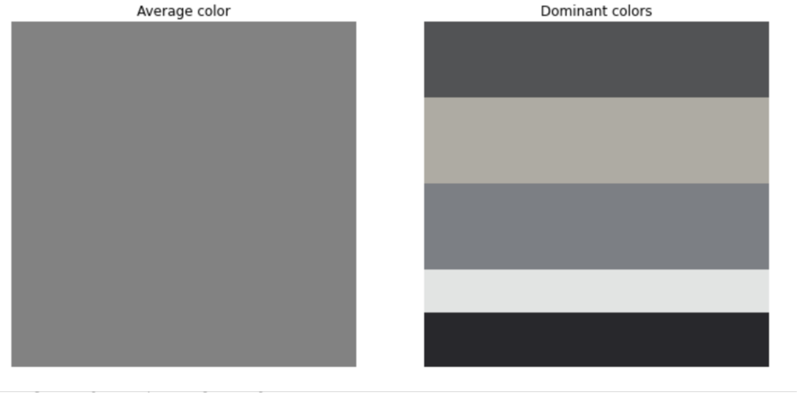
- The data consists of 10 classes of images: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks.
- Each of the images are 32x32 color pixels with 6,000 images per class.

Exploratory Data Analysis (EDA)

- Validated the normality of the distribution among the data set
- Plotted images in order to see their output
- Measured the most used colors of a given image as well as the average

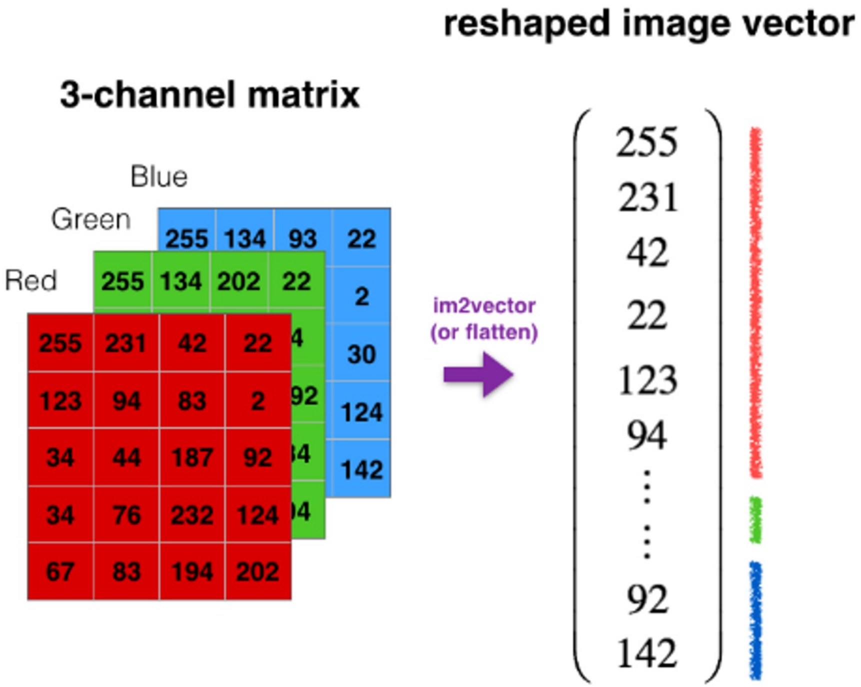






Data Pre-processing

- Binary classification:
  - Animals: bird, cat, deer, dog, frog, horse
  - Objects: airplane, automobile, ship, truck
- Flatten color images
  - RGB: 32 x 32 x 3 → 1 x 3,072
  - Couple horizontal and vertical data
- Split data
  - ⅔ training
  - ⅙ testing
  - ⅙ validation
- Normalize Data
  - Mean = 0, Standard deviation = 1

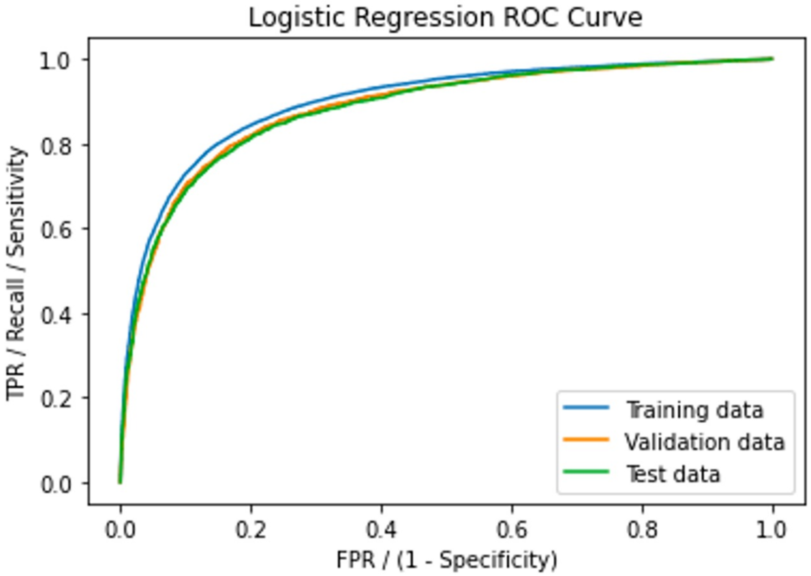


255	231	42	22
123	94	83	2
34	44	187	92
34	76	232	124
67	83	194	202

Model Analysis

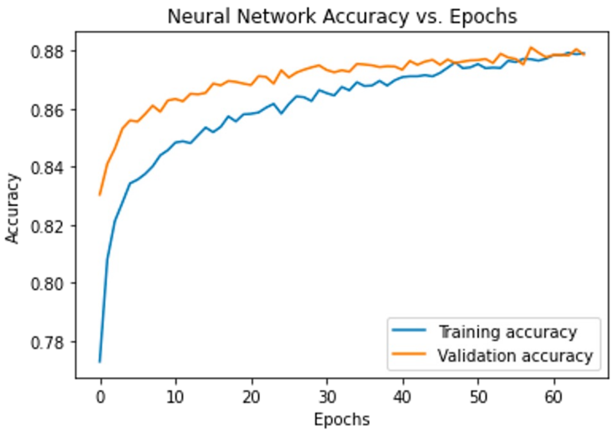
Initial Model - Logistic Regression (LR)

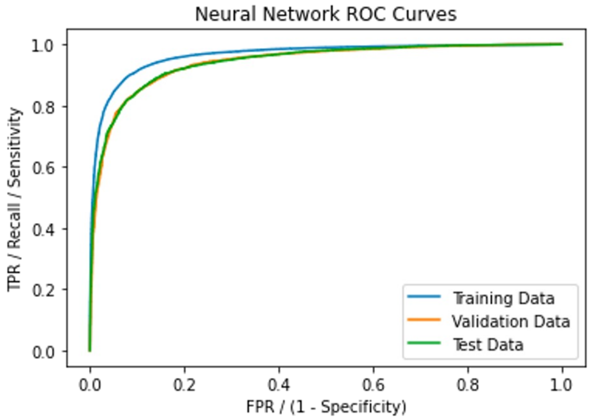
- Classification model using conditional probability
- Pros:
  - Simple to implement
  - Efficient to train
  - Understandable interpretations
- Cons:
  - Can overfit in high dimensional datasets
  - Does not support multicollinear data
- Results:
  - **Test Accuracy = 81.6%**



Optimization Model - Neural Network (NN)

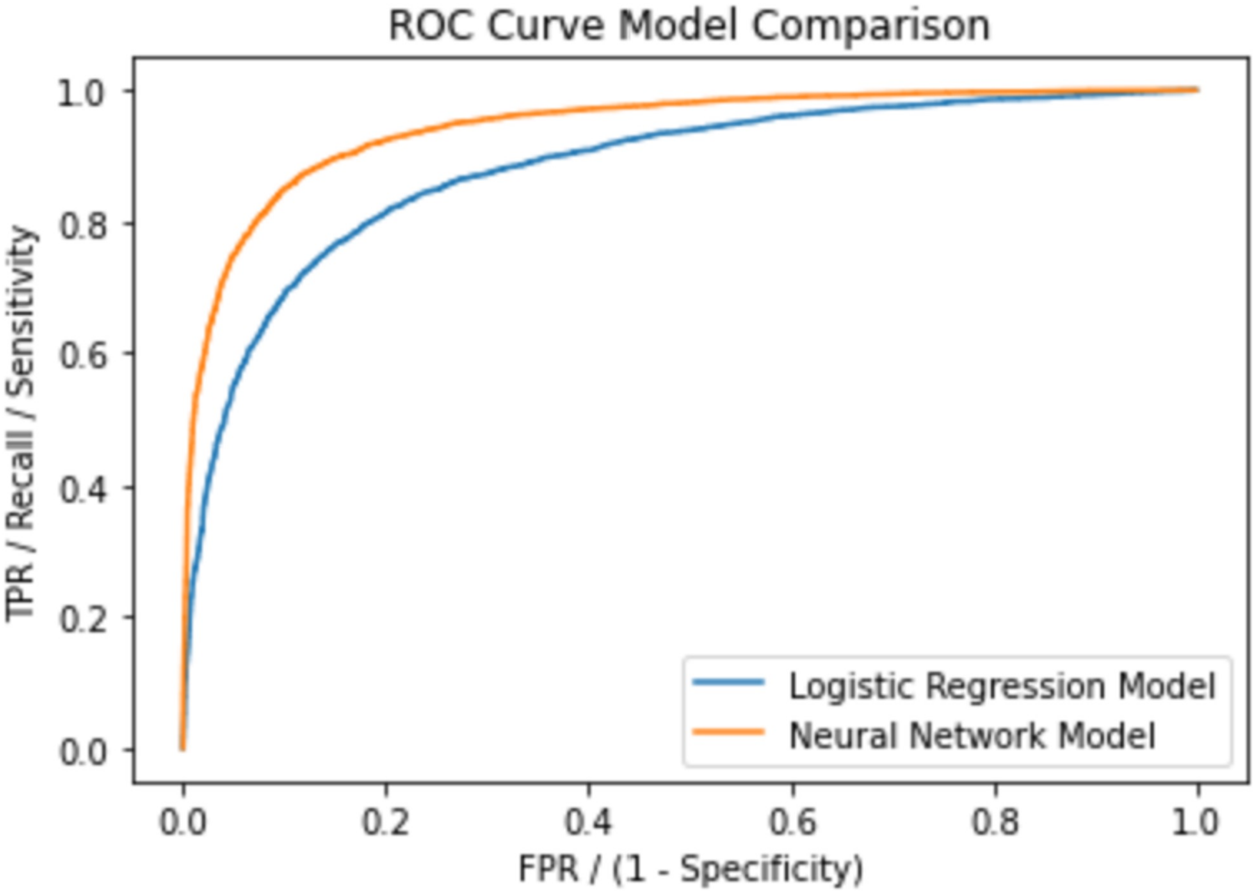
- Neural Network
  - Method in AI that teaches computers to process data in a way that is inspired by the human brain.
- Model Features
  - 1 Hidden Layer
    - 29 Neurons
    - Rectified Linear Unit (ReLU) Activation Function
  - Dropout Layer
    - 36% Dropout Rate
  - Output Layer
    - 1 Neuron
    - Sigmoid Activation Function
- Training Features
  - Stochastic Gradient Descent (SGD) Optimizer
  - Binary Cross Entropy Loss Function
  - Epochs = 65
  - Batch Size = 200
- Results
  - Test Accuracy = 87.8%





Compare Model Performances

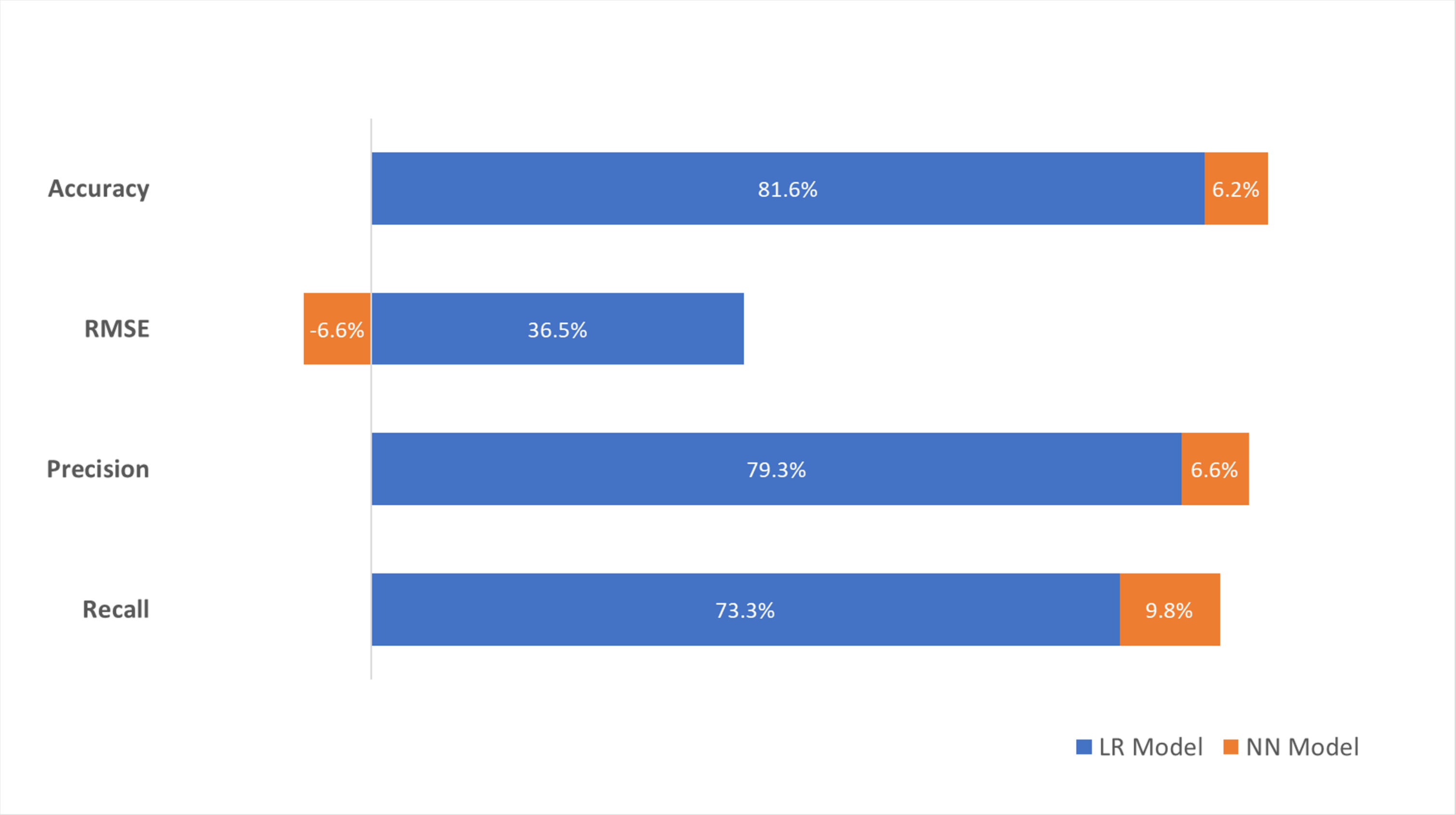
Note: We aim to achieve an AUC Score as closest to 1 as possible.



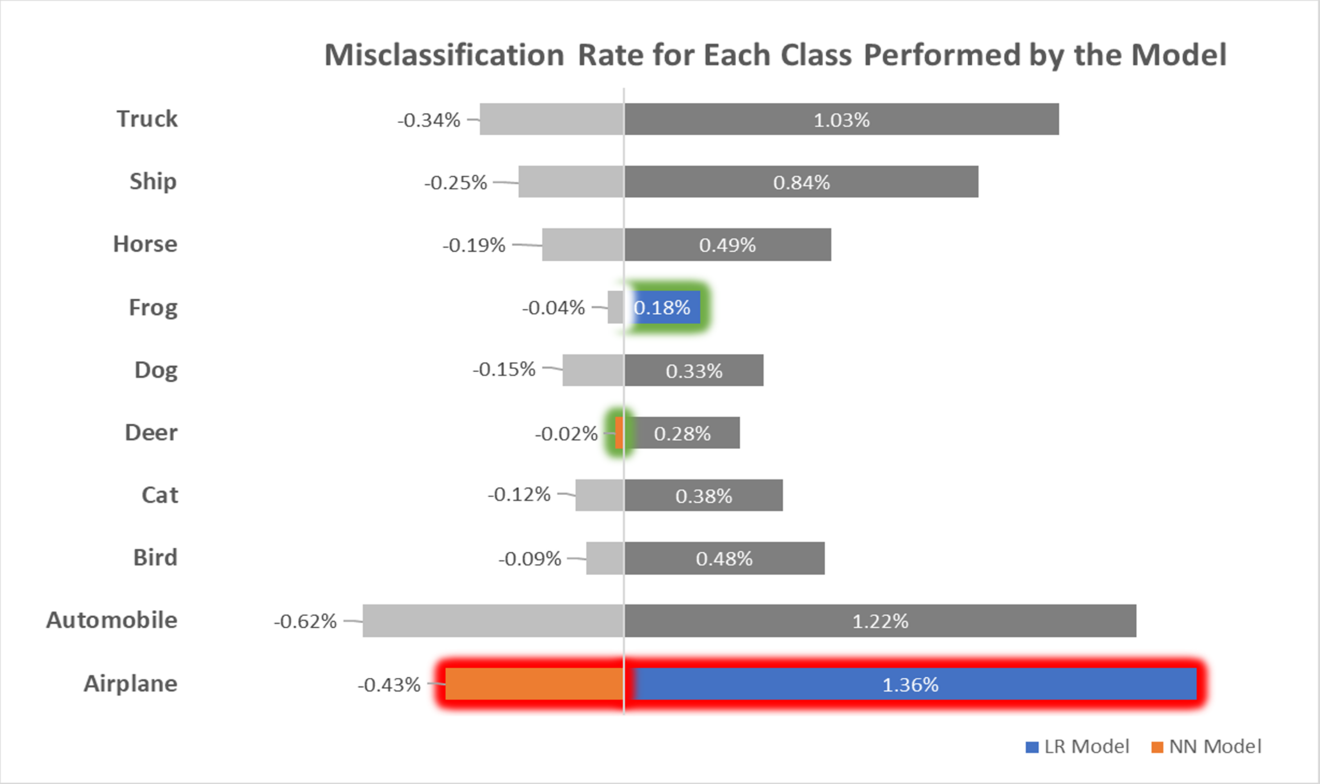
LR Model **AUC** Score: **0.803**

NN Model **AUC** Score: **0.942**

The model performance for the neural network (NN) had outperformed logistic regression (LR) in all model metric criticas.



The neural network model misclassification rate was lower than the logistic regression model for every image classification, where their overall rates were 12.2% and 18.4%, respectively.



Conclusion

- Our team recommends the optimization model for effectively identifying the appropriate image
- Overall model performance were better with the neural network vs the logistic regression model
  - E.g., Accuracy, RMSE, Precision, Recall, and AUC Score
- Misclassified Classes
  - Airplanes (most for both models)
  - Deer (least for NN model)
  - Frog (least for LR model)
  - Objects > living-things

References

1. Scott, E., January 18, *2018. CAPTCHAs Have an 8% Failure Rate, and 29% if Case Sensitive*, Baymard Institute
    a. https://baymard.com/blog/captchas-in-checkout
2. N, B. 2022, October 3-10, 2022. [Lecture recording]. Notre Dame University.
    a. Dimensionality Reduction in Images
        i. https://notredame.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c1370d42-4327-48e0-aede-af2400100d7c
    b. Deep Learning
        i. https://notredame.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0473db1f-6a1a-433b-b289-af2b001518b1
3. Koehren, W., September 15, 2022. Use Precision and Recall to Evaluate Your Classification Model When Accurately Isn’t Enough, Built In
    a. https://builtin.com/data-science/precision-and-recall

Thank you for listening!