David M. Arnold

Data Science – Winter 2021 Cohort

04/06/2021

USING MACHINE LEARNING TO DETERMINE WHAT MAKES A REDDIT MEME POPULAR

Overview

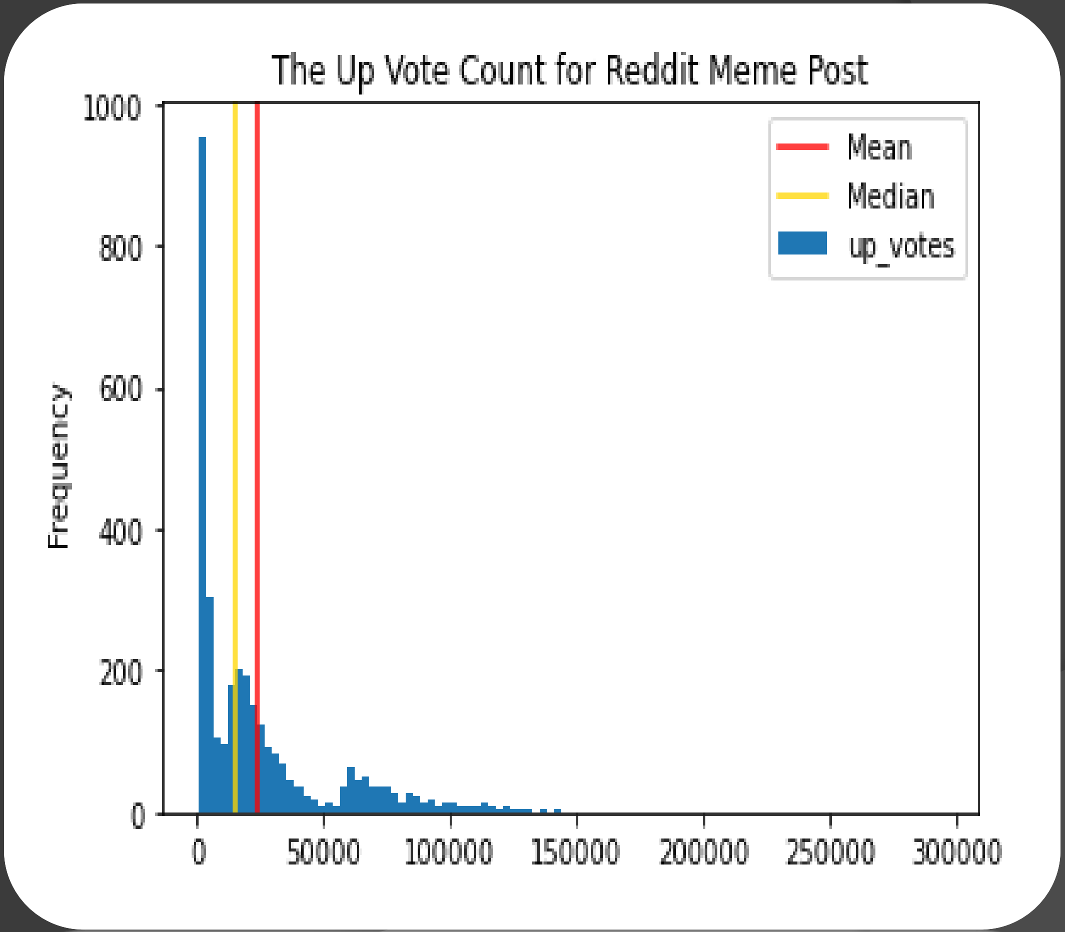
- Project Introduction / Background
- Target / Features
- Machine Learning Model Results
- Image Classification Results
- Summary & Recommendations

Project Introduction /Background

- What Makes a Popular Meme On Reddit?
  - Content
    - Sentiment / Feeling
  - User Submitting
  - Date Posted
  - Title / Description

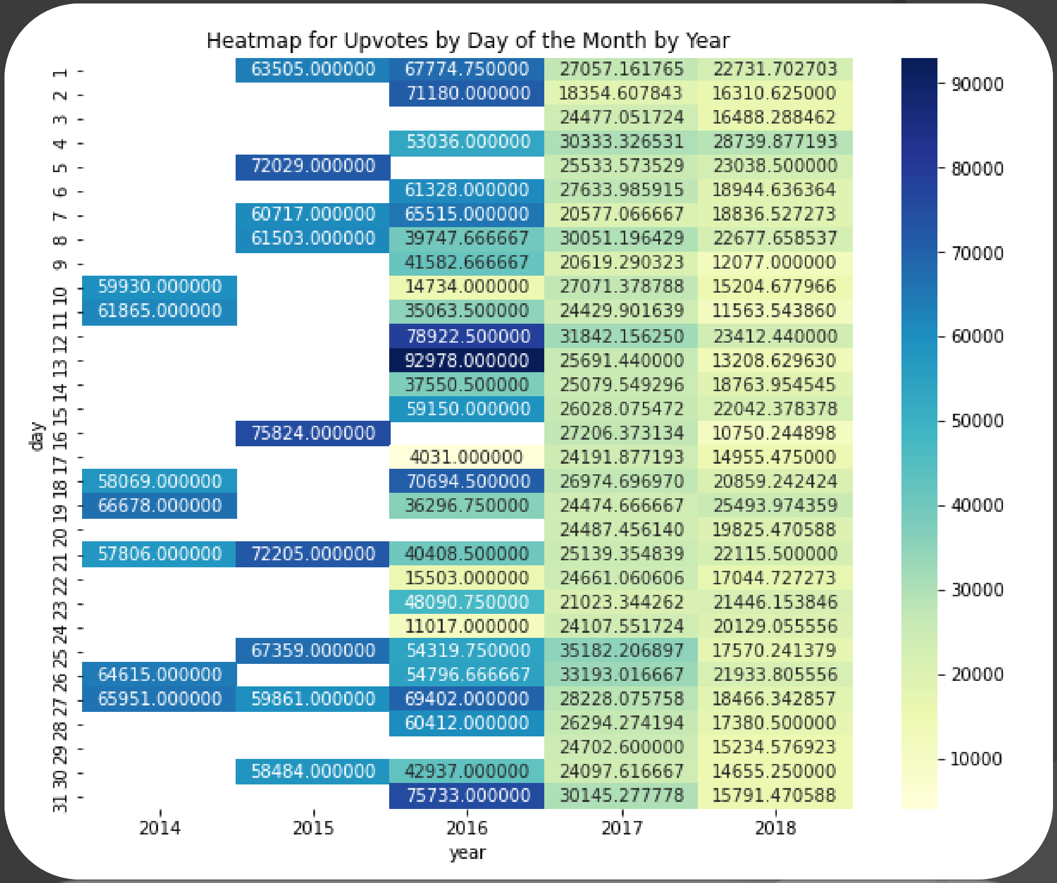
Target – Upvotes

- Upvotes
- Binary Classification
  - Good Meme
  - Over 24,000 votes
  - Bad Meme
  - Under 24,000 votes



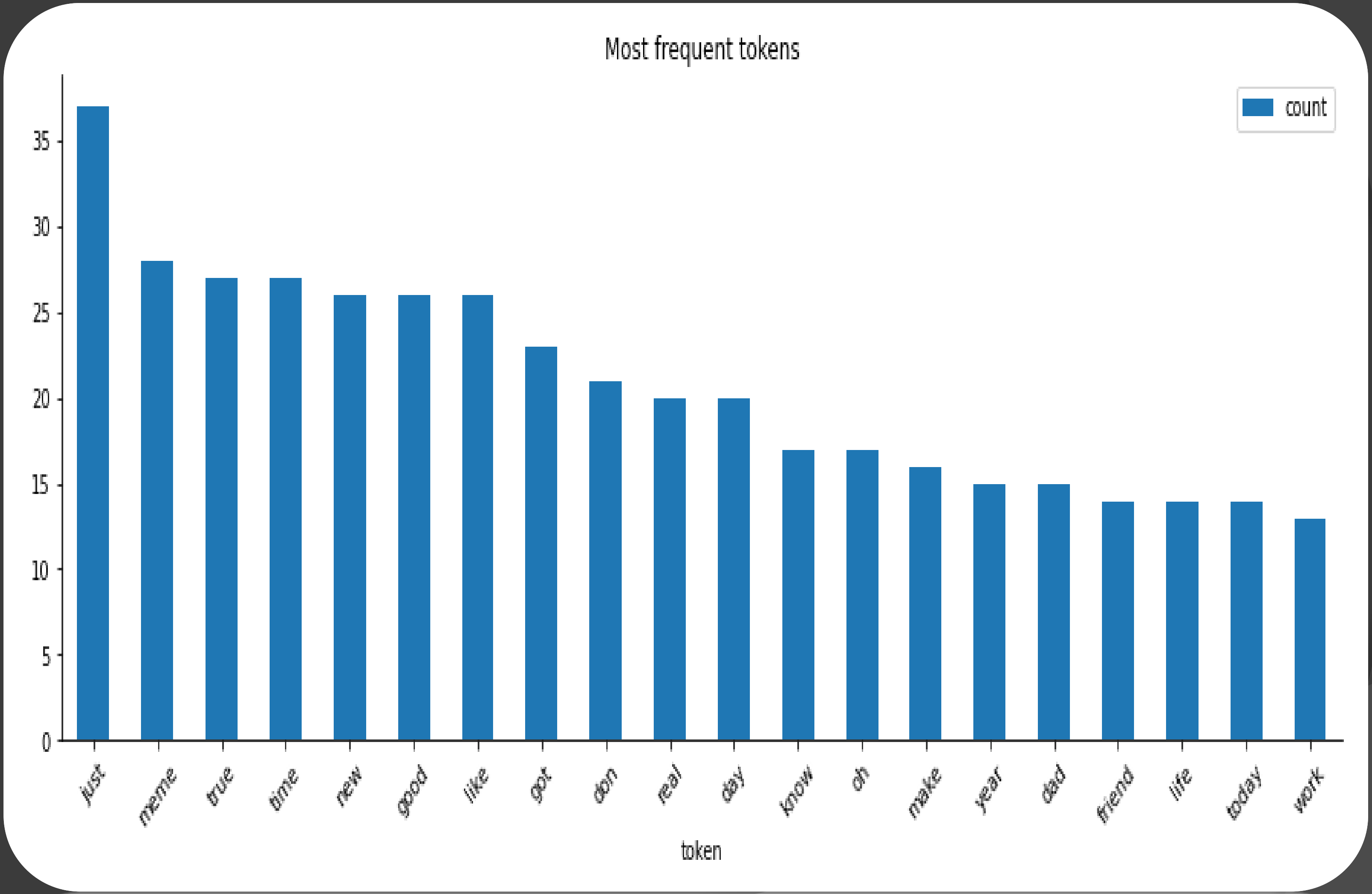
Features – Date Posted

- Fewer Posts / More Votes
- Separate into
  - Year
  - Months
  - Weekdays
  - Weekends

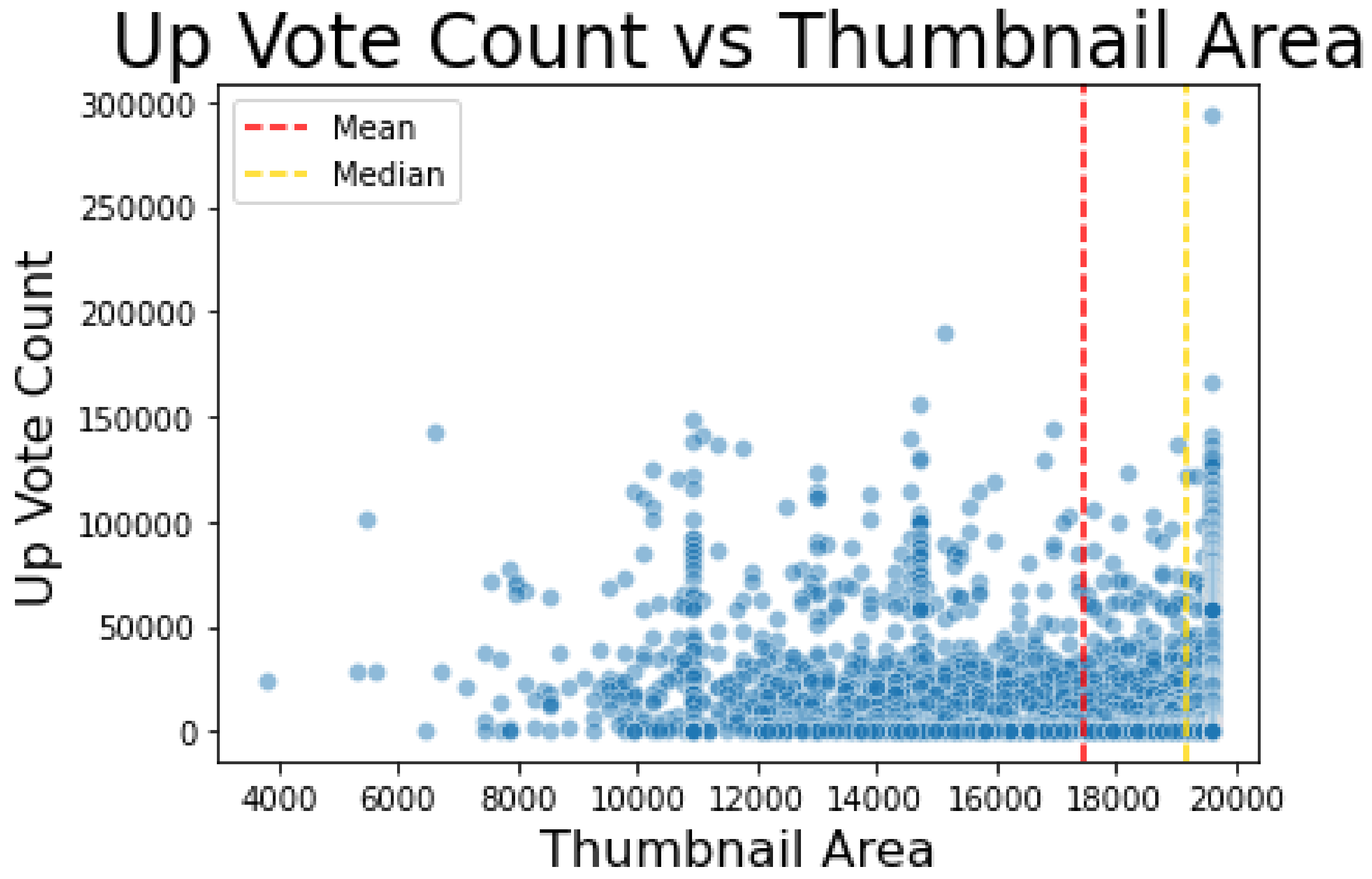


day	2014	2015	2016	2017	2018
1		63505.000000	67774.750000	27057.161765	22731.702703
2			71180.000000	18354.607843	16310.625000
3				24477.051724	16488.288462
4			53036.000000	30333.326531	28739.877193
5		72029.000000		25533.573529	23038.500000
6			61328.000000	27633.985915	18944.636364
7		60717.000000	65515.000000	20577.066667	18836.527273
8		61503.000000	39747.666667	30051.196429	22677.658537
9			41582.666667	20619.290323	12077.000000
10	59930.000000		14734.000000	27071.378788	15204.677966
11	61865.000000		35063.500000	24429.901639	11563.543860
12			78922.500000	31842.156250	23412.440000
13			92978.000000	25691.440000	13208.629630
14			37550.500000	25079.549296	18763.954545
15			59150.000000	26028.075472	22042.378378
16		75824.000000		27206.373134	10750.244898
17			4031.000000	24191.877193	14955.475000
18	58069.000000		70694.500000	26974.696970	20859.242424
19	66678.000000		36296.750000	24474.666667	25493.974359
20				24487.456140	19825.470588
21	57806.000000	72205.000000	40408.500000	25139.354839	22115.500000
22			15503.000000	24661.060606	17044.727273
23			48090.750000	21023.344262	21446.153846
24			11017.000000	24107.551724	20129.055556
25		67359.000000	54319.750000	35182.206897	17570.241379
26	64615.000000		54796.666667	33193.016667	21933.805556
27	65951.000000	59861.000000	69402.000000	28228.075758	18466.342857
28			60412.000000	26294.274194	17380.500000
29				24702.600000	15234.576923
30		58484.000000	42937.000000	24097.616667	14655.250000
31			75733.000000	30145.277778	15791.470588

Features – Title Column

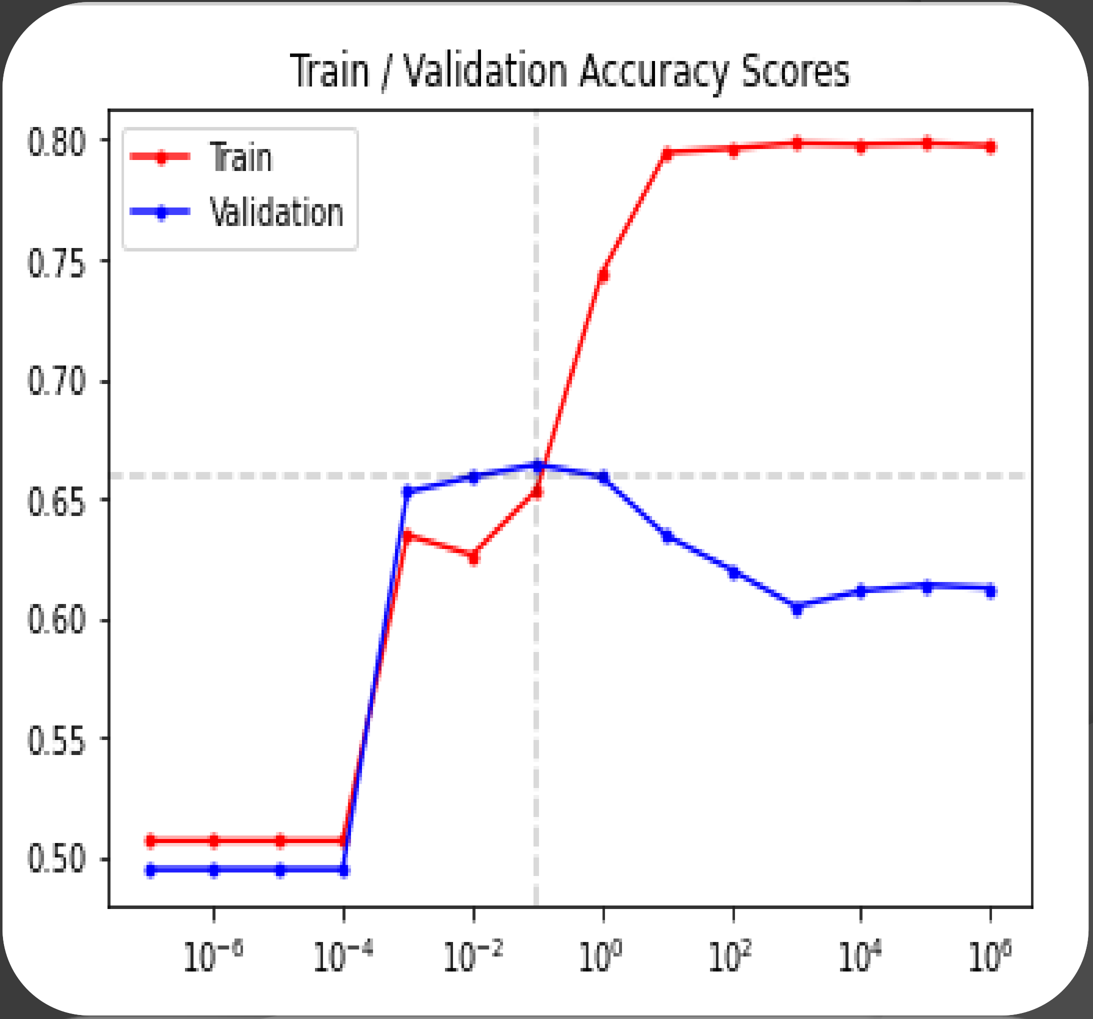


Features – Thumbnail Area



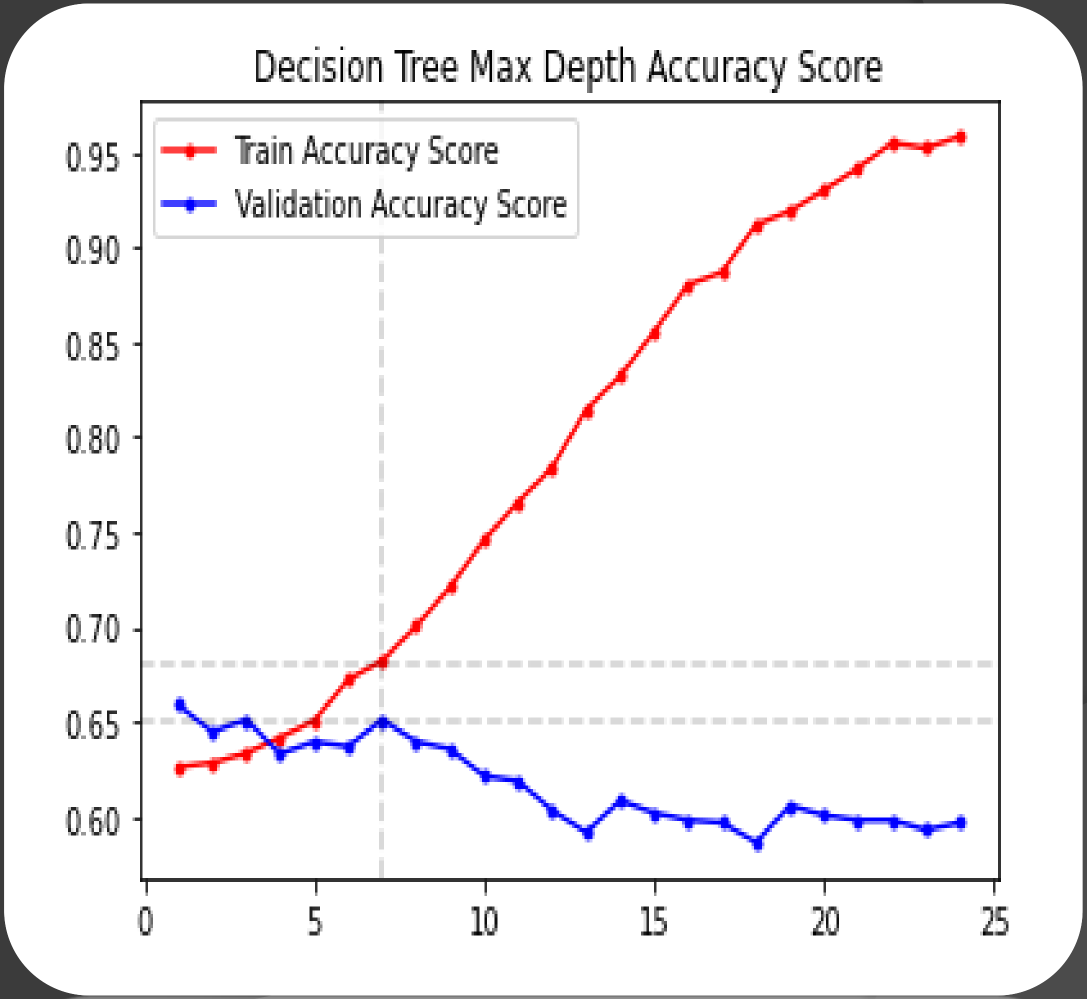
ML Modeling – Logistic Regression

- “C” value – 0.1
- MinMax Scaler
- PCA 10 features
- Accuracy Score
  - Train 65.3%
  - Validation 65%
  - Test 66.4%



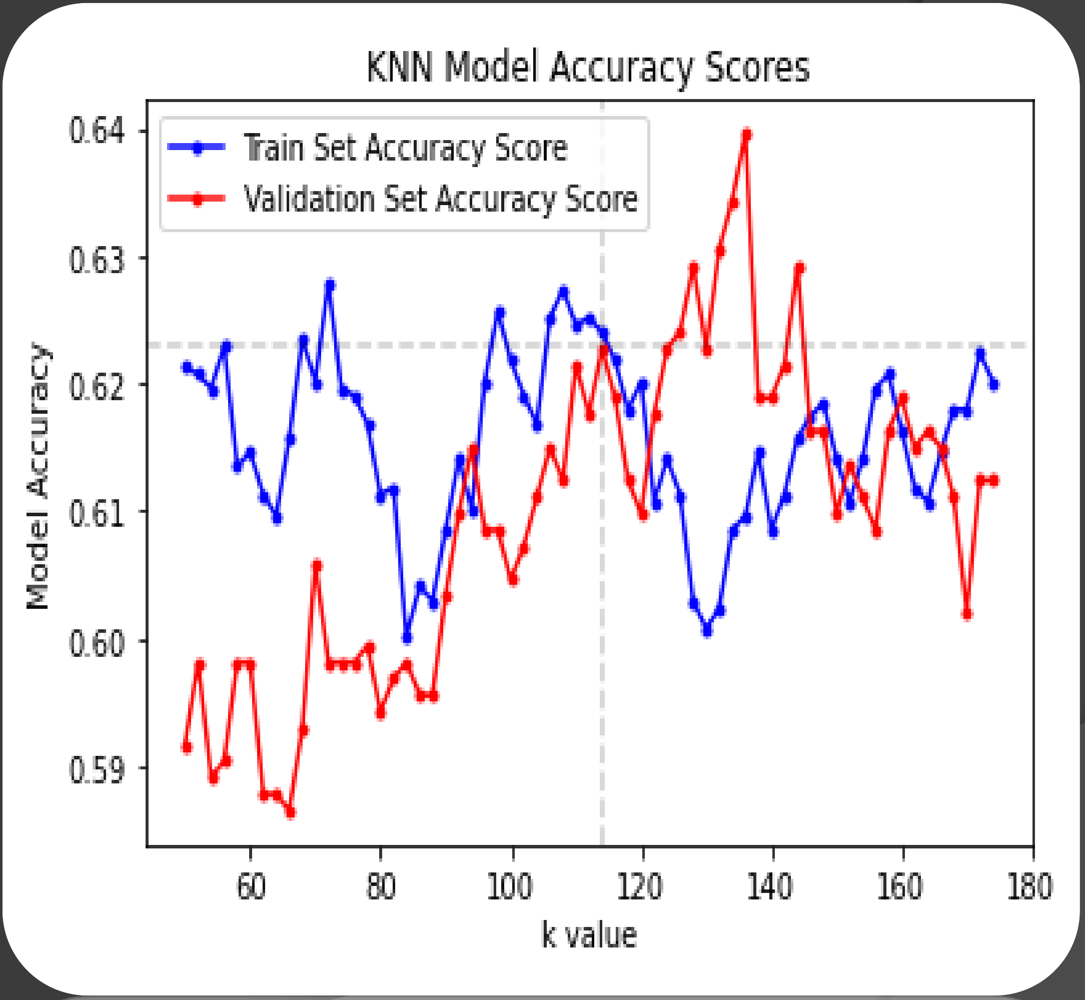
ML Modeling – Decision Trees

- Max Depth = 7
- No scaling/PCA
- Accuracy Score
  - Train 68.2%
  - Validation 65%
  - Test 65.2%



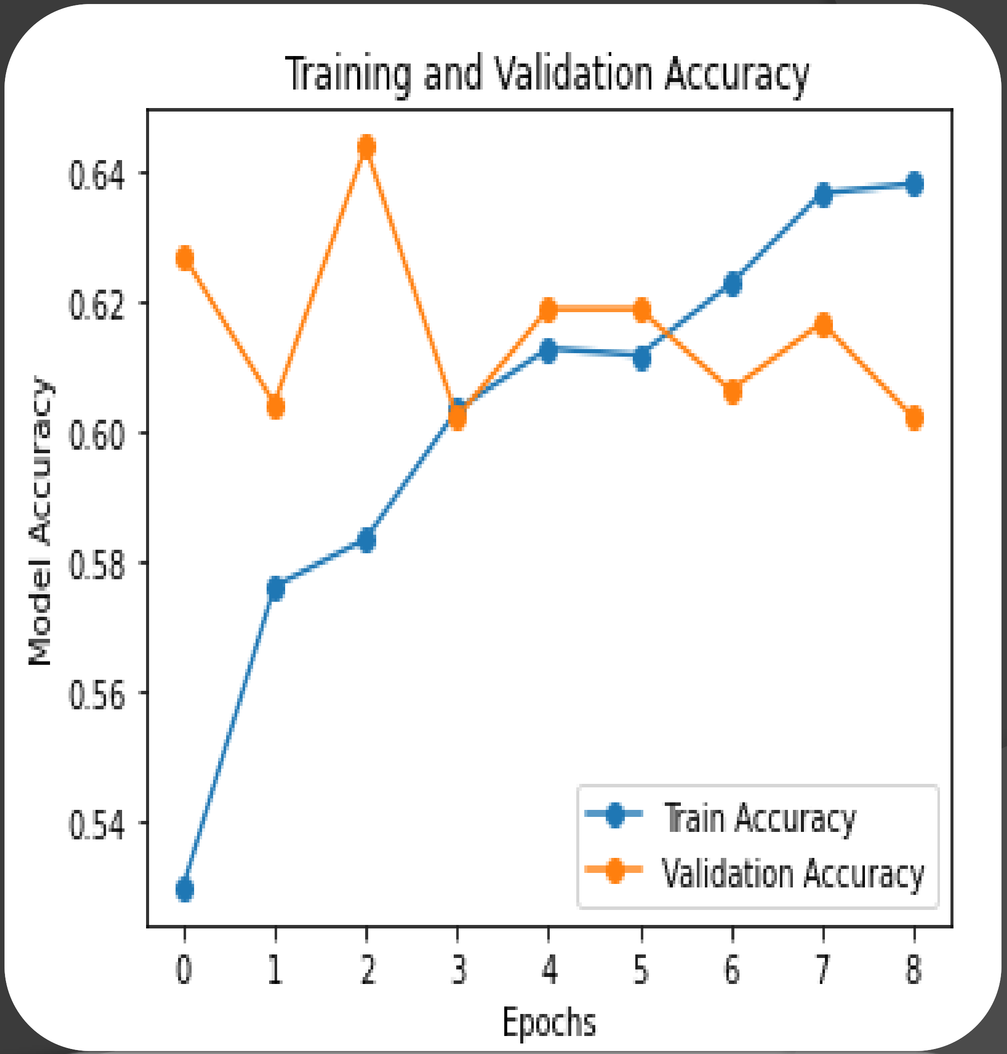
ML Modeling – K-Nearest Neighbors

- "k" value – 66
- MinMax Scaler
- Accuracy Score
  - Train 61.6%
  - Validation 62%
  - Test 59.1%



ML Modeling – Convolutional Neural Network (CNN)

- Classification
  - Good / Bad
- Pre-Trained Model
  - VGG16
- Pixel Size
  - 64 x 64
  - 256 x 256
- Accuracy
  - 62-65%



Summary

- ML / Neural Networks
  - 59 – 68% Test Accuracy
  - Decision Trees Best Results
- Need Confusion Matrices
- Sample / Predict Images
- Add More Datapoints
- Prediction on Newer Memes
- Share on Kaggle