GGE

Good Games For Everyone

Jerome Sim, Joey Kang, Tay Yi Li, Adriel Chen

TABLE OF CONTENTS

01

-

02

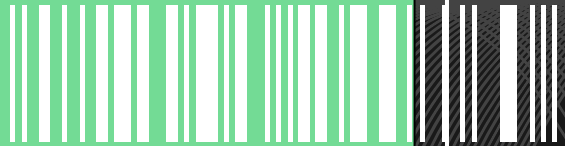
03

04

05

06

01



BACKGROUND

About GGE and our data science goals

ABOUT US

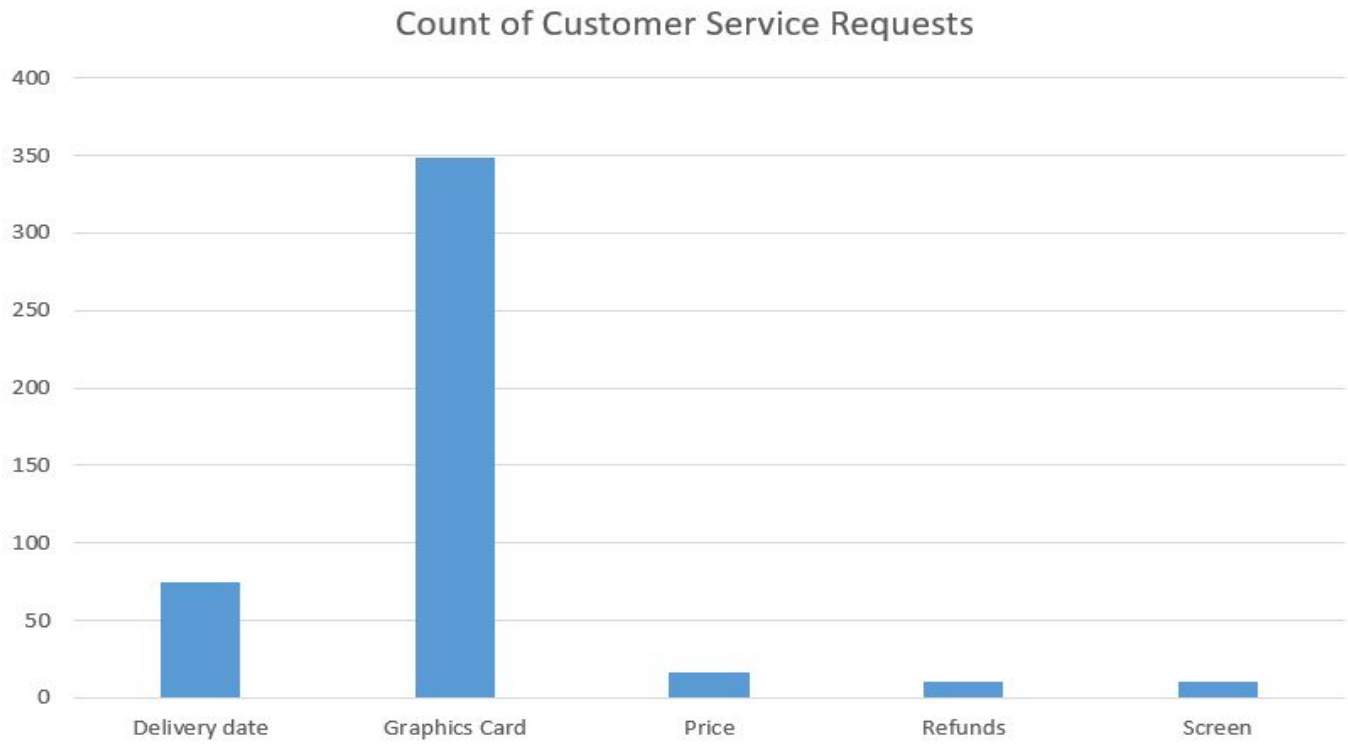
Good Games for Everyone is a PC building company, specialising in constructing high-performance gaming terminals for our customers. Adriel, Joey, and Yi Li make up their Data Science Division.

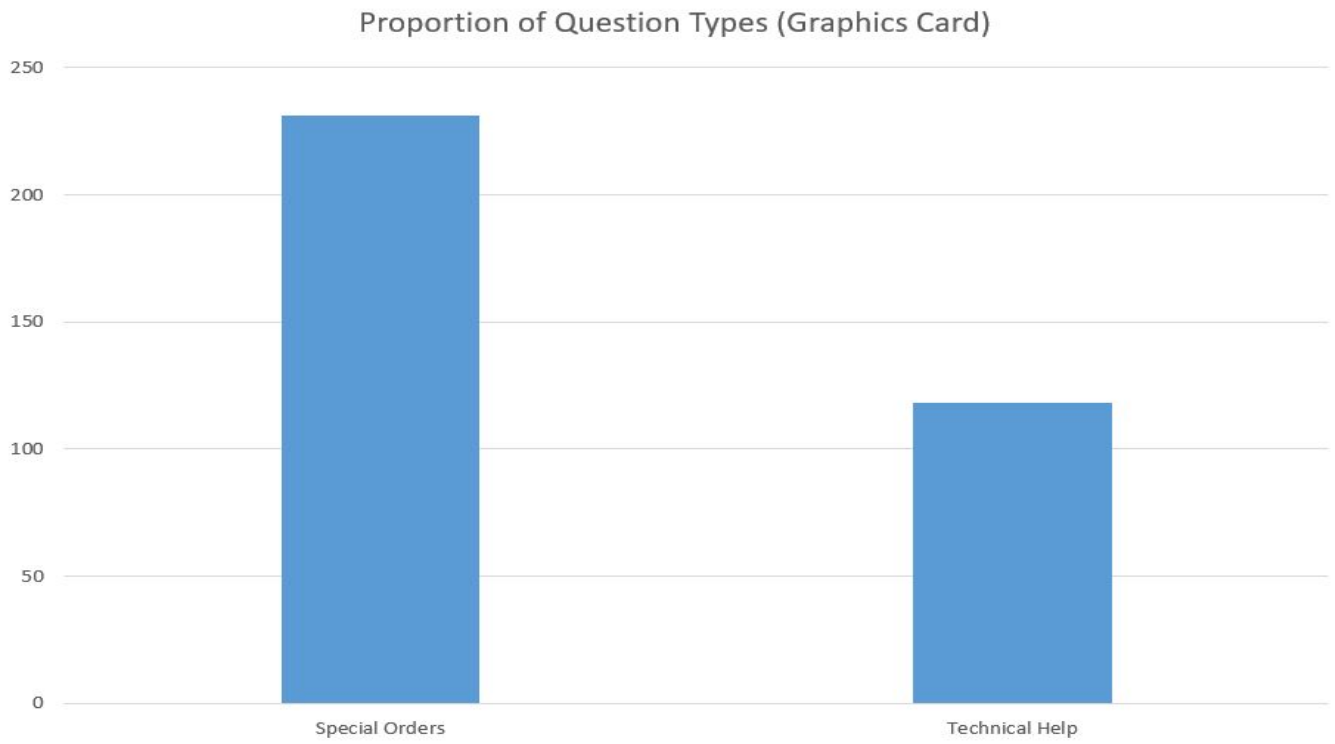
Div Head,
DS
Adriel Chen

Team Lead,
DS
Joey Kang

Team Lead,
DS
Tay Yi Li

Chief
Hygienist
Jerome Sim

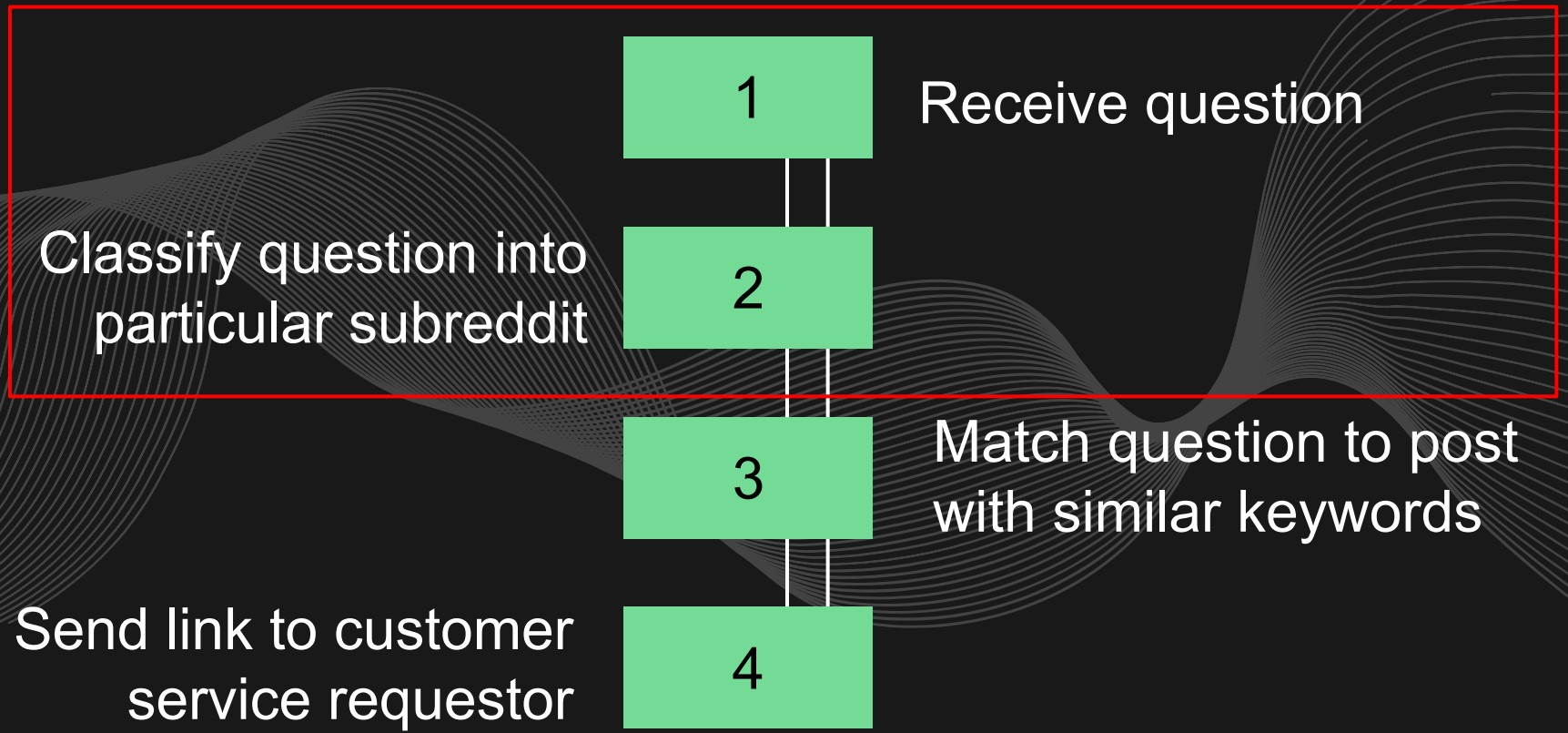

Count of Customer Service Requests
400
350
300
250
200
150
100
50
0
Delivery date
Graphics Card
Price
Refunds
Screen


Proportion of Question Types (Graphics Card)
250
200
150
100
50
0
Special Orders
Technical Help

Problem Statement

Management tasked the Data Science and Software Engineering teams to find a way to field after-hours requests for technical help regarding graphics cards.

The Big Picture



Goals

Aim
To build a classifier that identifies keywords for the bot to determine relevant subreddits accurately.

KPIs:

- Sensitivity
- Specificity
- Accuracy

01

-

02

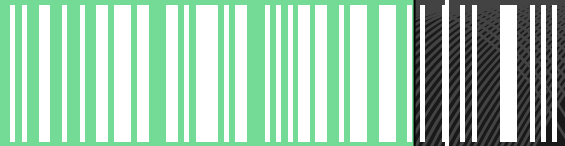
03

04

05

06

02



DATA PROCESSING

Web-scraping, Data Cleaning and Exploratory Data Analysis

WEB-SCRAPING

PushShift API to scrape from Reddit's r/Nvidia and r/AMD.

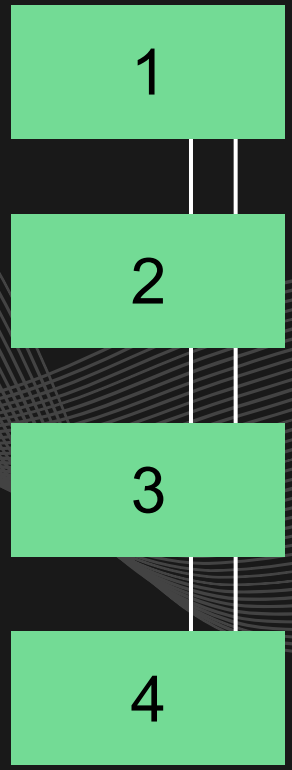
Data:

- 10000 rows each
- Variables: Title, Post's text, Author, Subreddit, Time Stamp, Number of Comments, Number of Crossposts, Post's Score, Total Awards Received by Post.

DATA CLEANING

- Kept useful columns
- Removed duplicates, deleted posts, and posts containing photos/videos
- Combined title and Post's text columns
- Concatenated Nvidia and AMD dataframes

PRE-PROCESSING TEXT



Removing Links, Punctuations

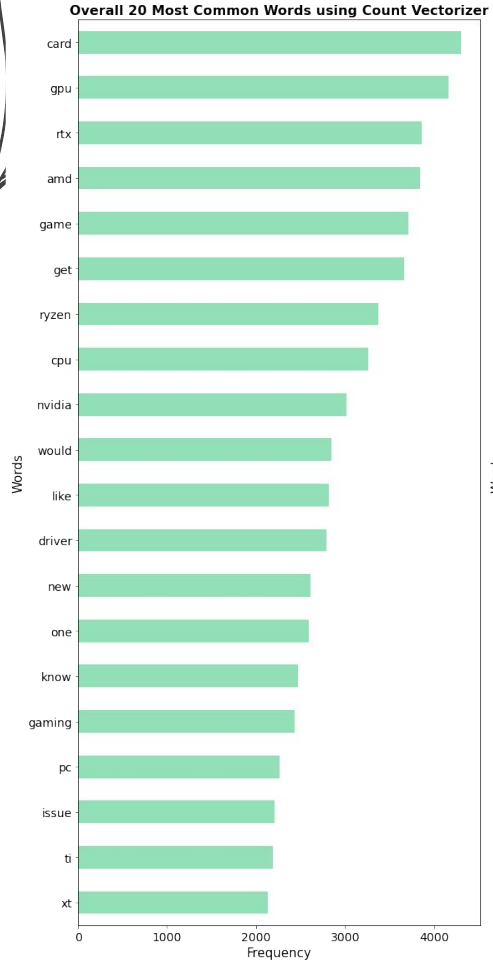
Removing stopwords from nltk package

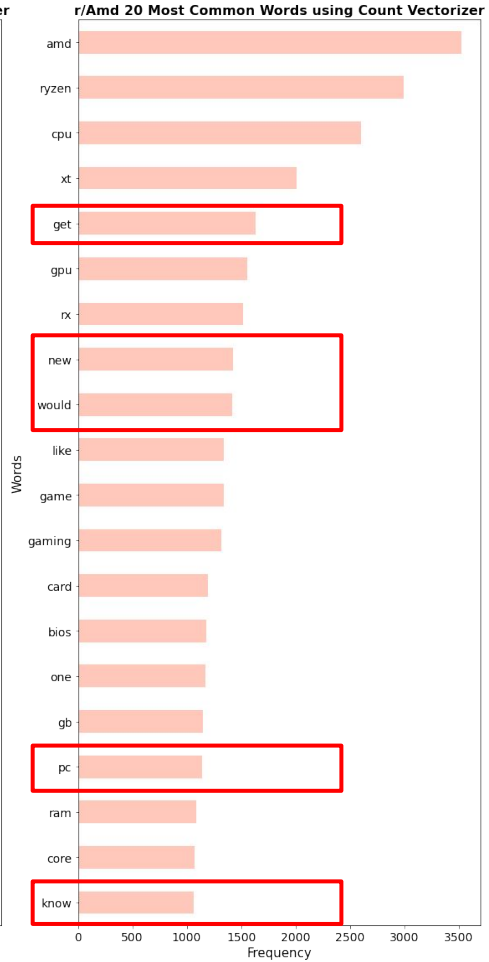
Removing additional stopwords identified through EDA

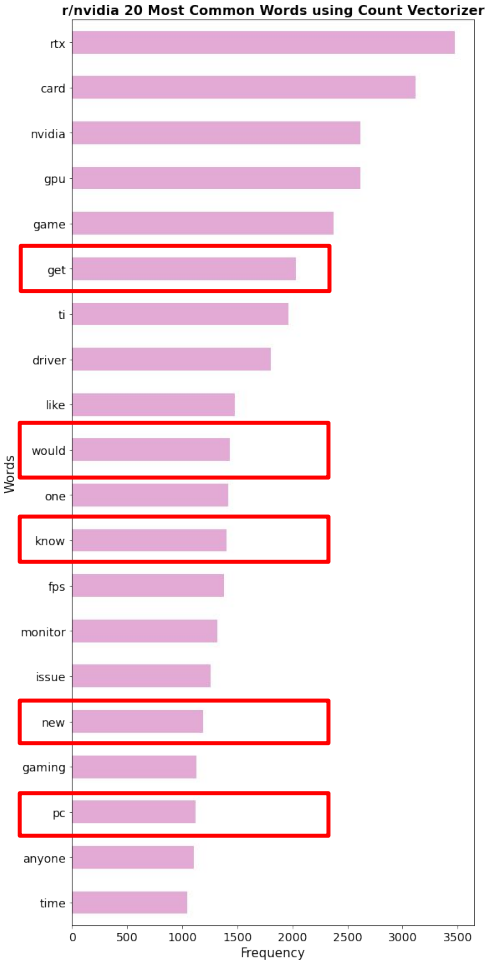
Lemmatization

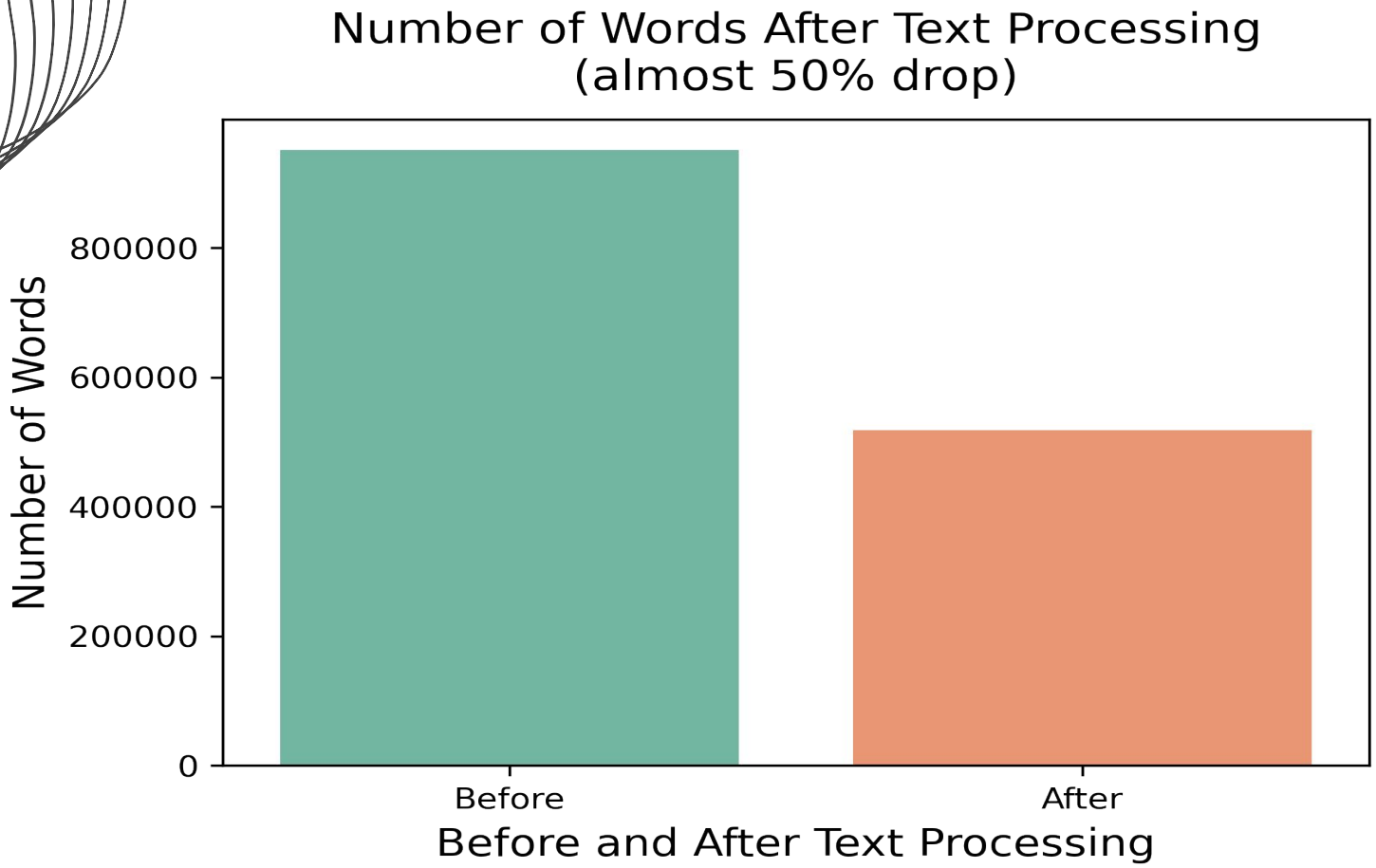
Removed generic words that appear in both subreddits:

- get
- new
- would
- pc
- know



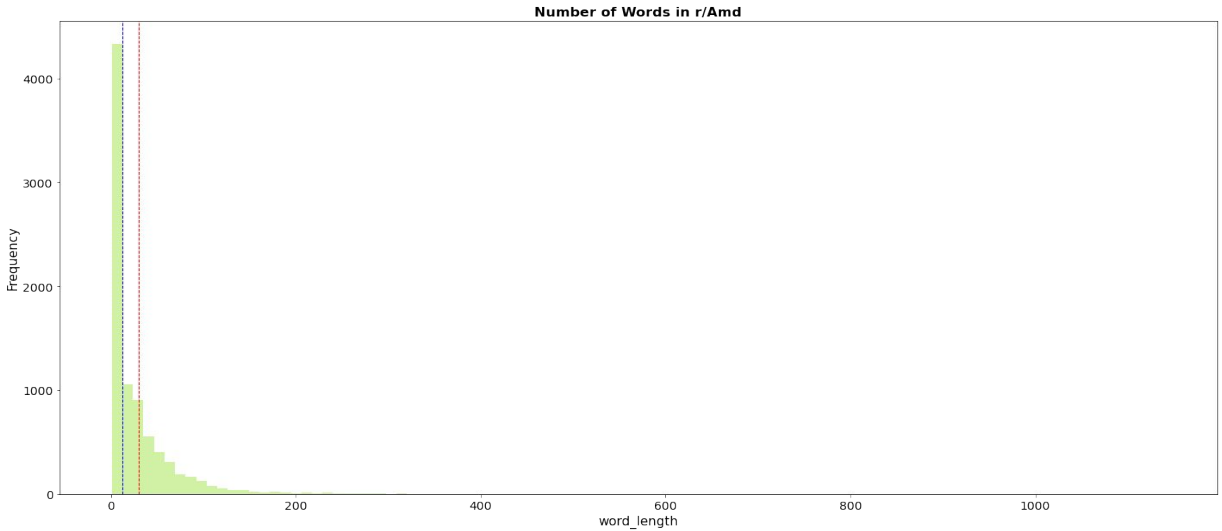


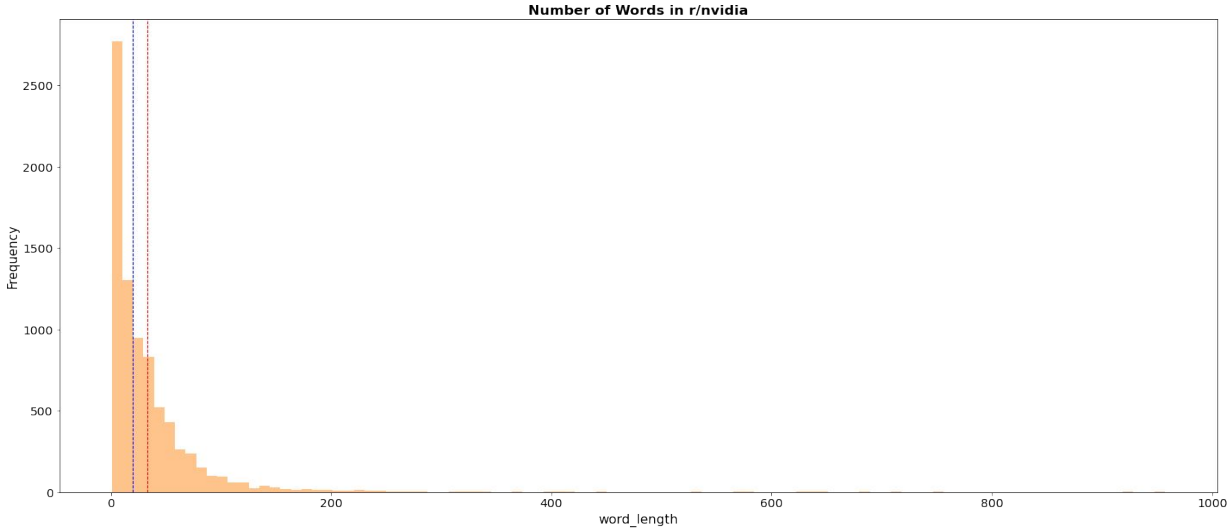



Number of Words After Text Processing
(almost 50% drop)
Number of Words
800000
600000
400000
200000
0
Before
After
Before and After Text Processing

- Distribution of word length is **positively skewed** in both subreddits

- Average word length for Amd (31) and Nvidia (33) is comparable





No correlation between any of the numerical columns with target variable (i.e. subreddit)

01

02

03

04

05

06

03

MODELLING

Model evaluation and analysis

MODELS EVALUATION

1

Decision Tree

1. Test Accuracy
2. AUC
3. Sensitivity & Specificity

2

Multinomial Naive Bayes

1. Test Accuracy
2. AUC
3. Sensitivity & Specificity

3

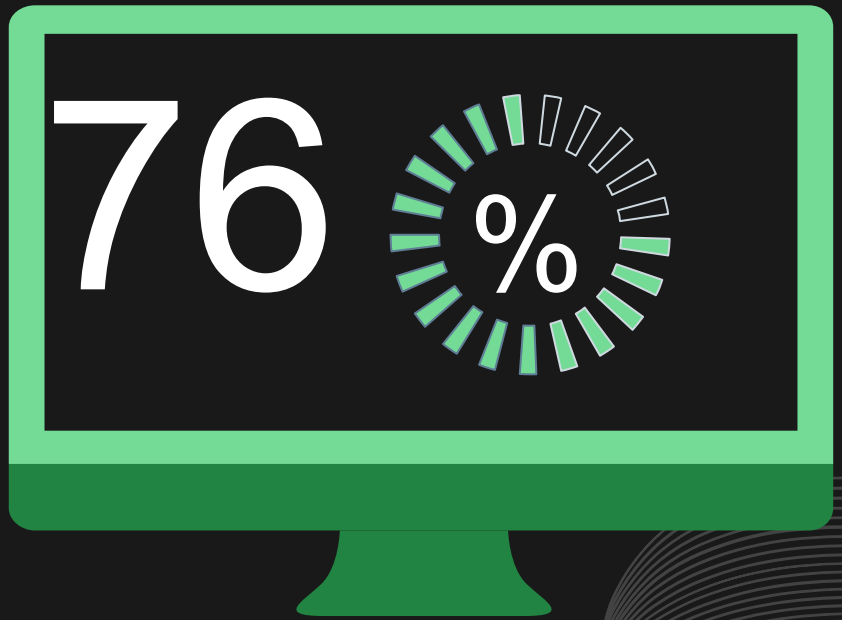
Logistic Regression

1. Test Accuracy
2. AUC
3. Sensitivity & Specificity

*Models use Tfidf Vectorizer

1. Decision Tree

TEST ACCURACY

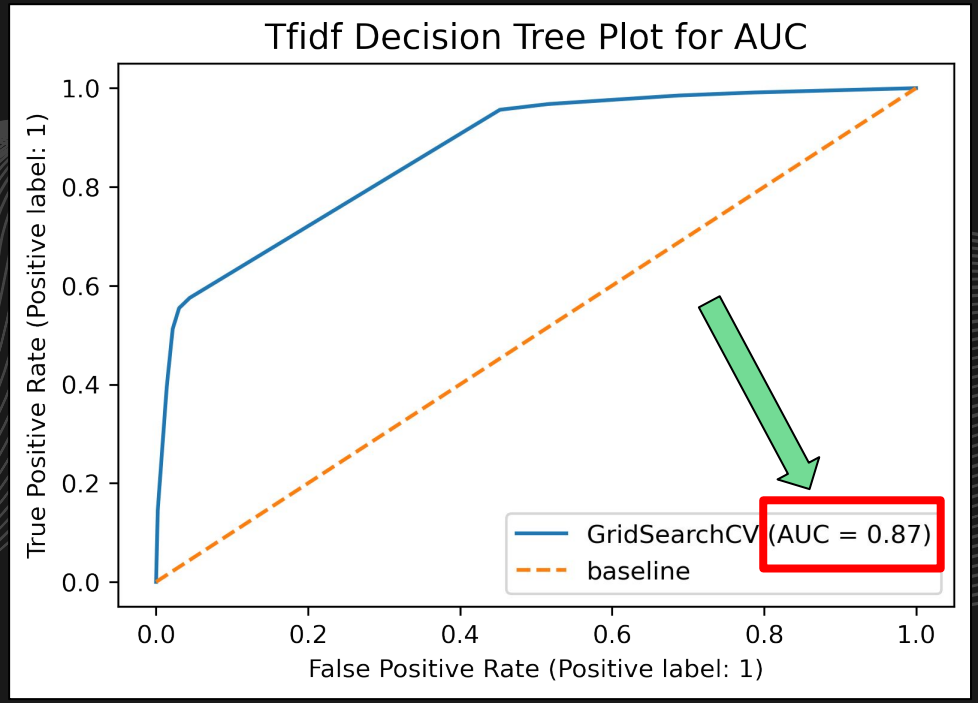


Overall ability to classify posts into the correct subreddit

(True Positives + True Negatives) / Total Predictions

1. Decision Tree

Area Under Curve



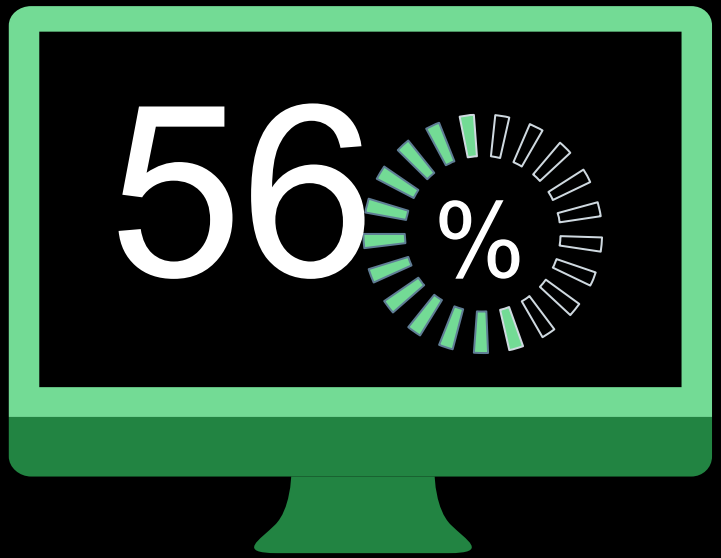
Good Model

Larger AUC

Better at distinguishing between Positives & Negatives

1. Decision Tree

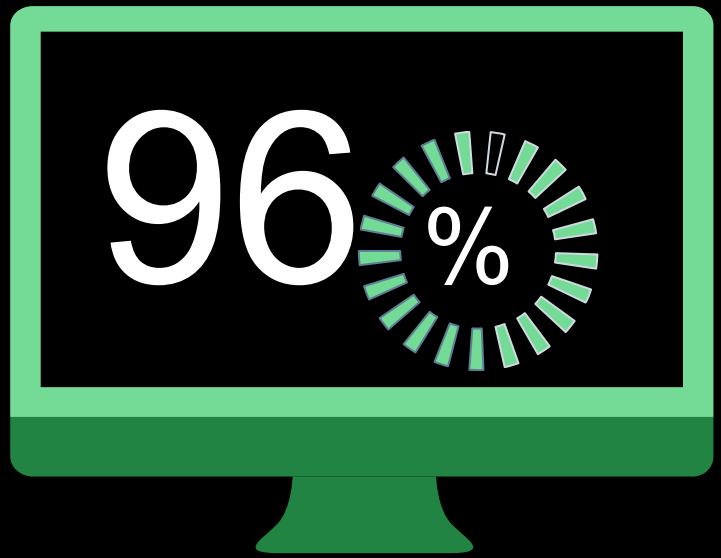
SENSITIVITY



Ratio of True Positives to All Actual Positives

Ability to correctly classify r/AMD posts

SPECIFICITY

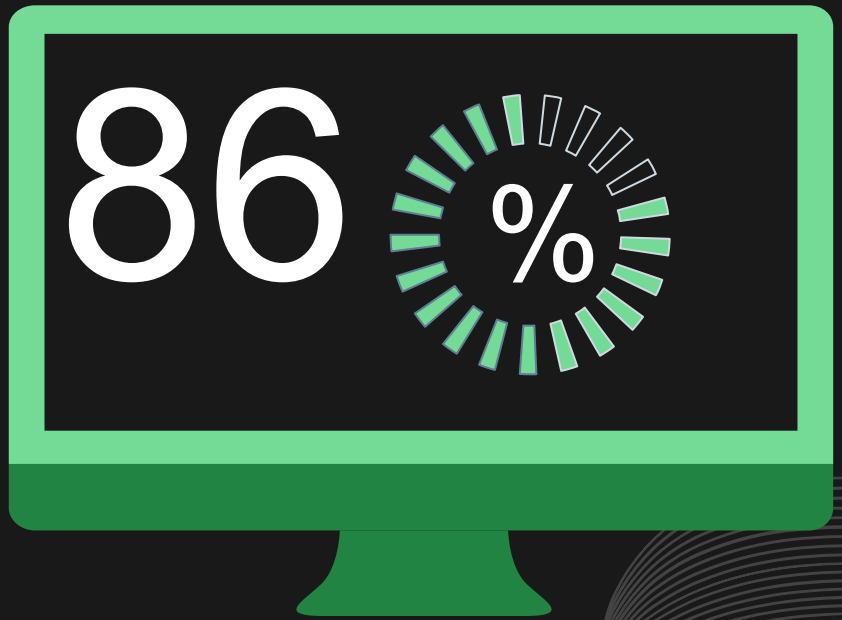


Ratio of True Negatives to All Actual Negatives

Ability to correctly classify r/NVIDIA posts

2. Multinomial Naive Bayes

TEST ACCURACY

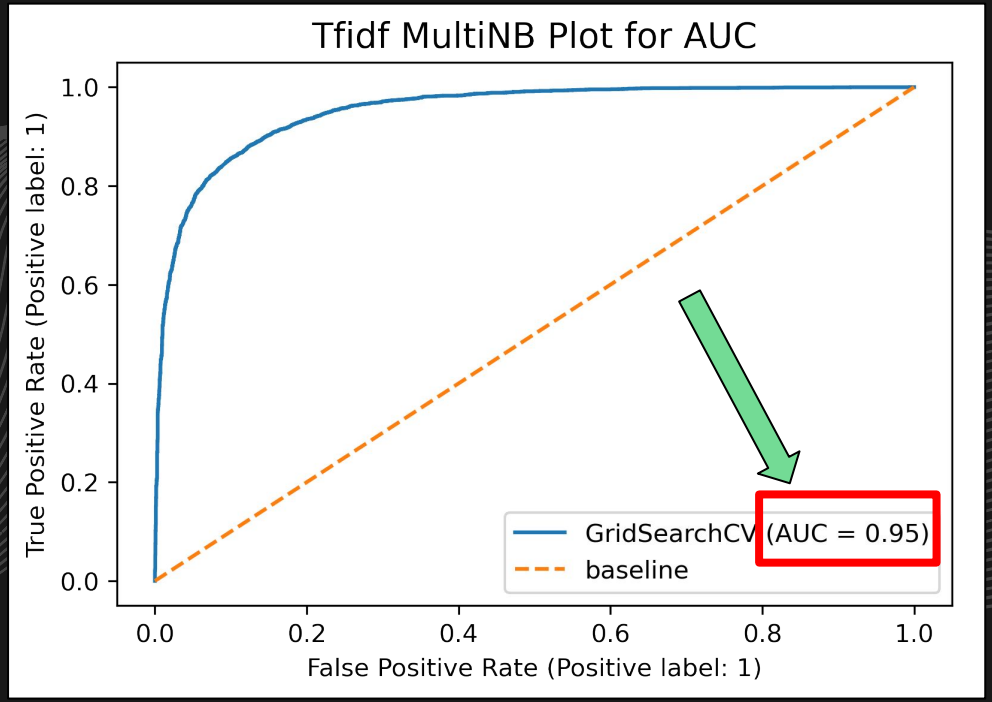


Overall ability to classify posts into the correct subreddit

(True Positives + True Negatives) / Total Predictions

2. Multinomial Naive Bayes

Area Under Curve



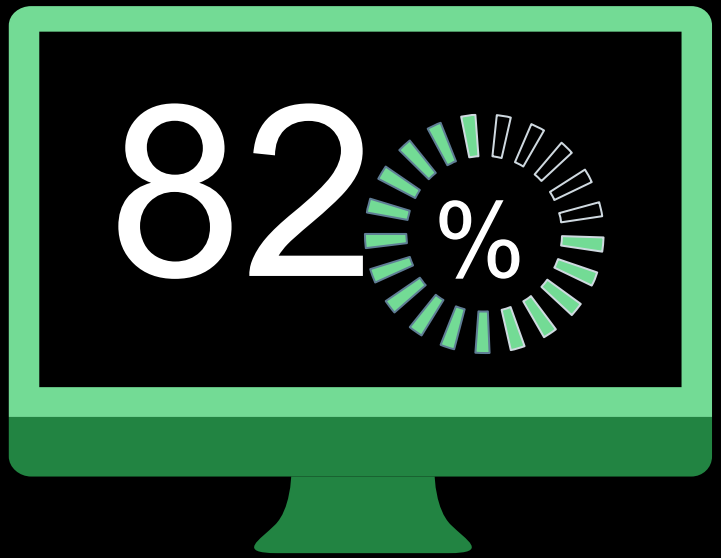
Good Model

Larger AUC

Better at distinguishing between Positives & Negatives

2. Multinomial Naive Bayes

SENSITIVITY



Ratio of True Positives to All Actual Positives

Ability to correctly classify r/AMD posts

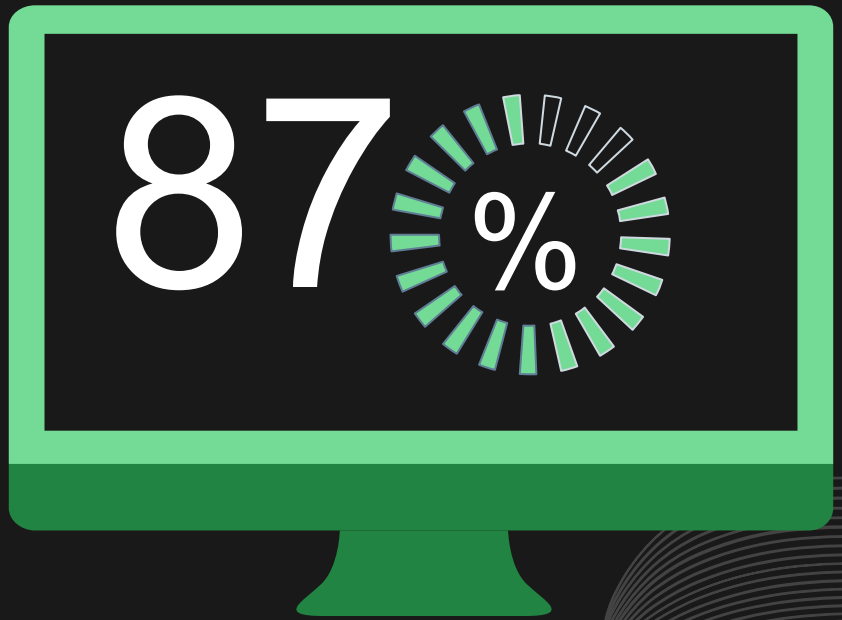
SPECIFICITY

Ratio of True Negatives to All Actual Negatives

Ability to correctly classify r/NVIDIA posts

3. Logistic Regression

TEST ACCURACY

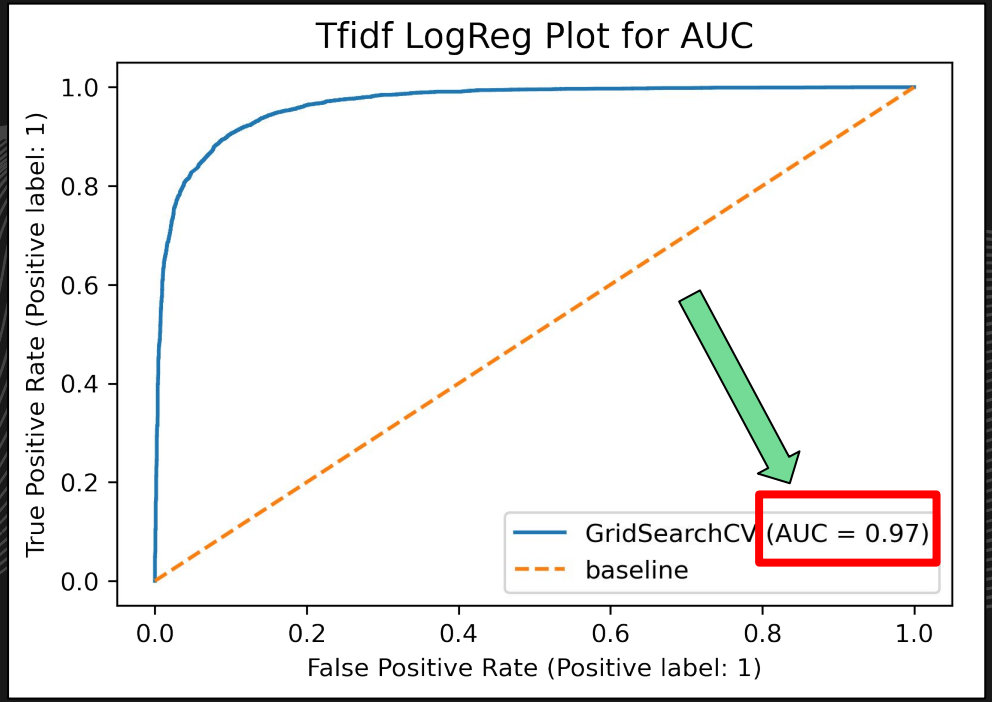


Overall ability to classify posts into the correct subreddit

(True Positives + True Negatives) / Total Predictions

3. Logistic Regression

Area Under Curve



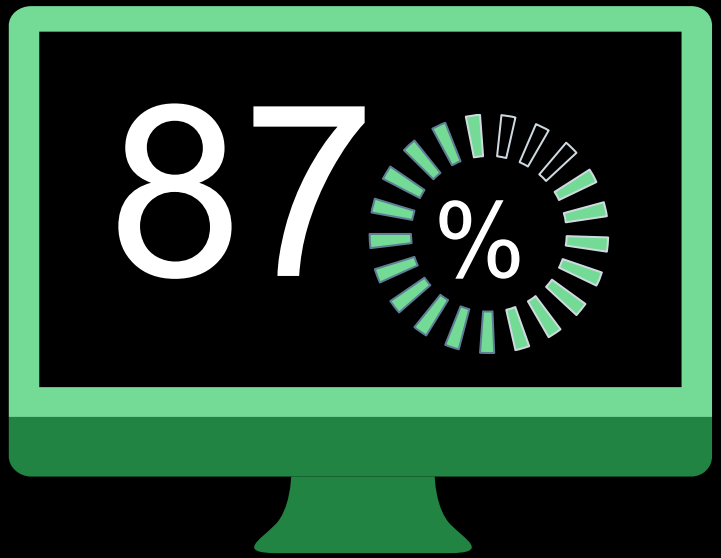
Good Model

Larger AUC

Better at distinguishing between Positives & Negatives

3. Logistic Regression

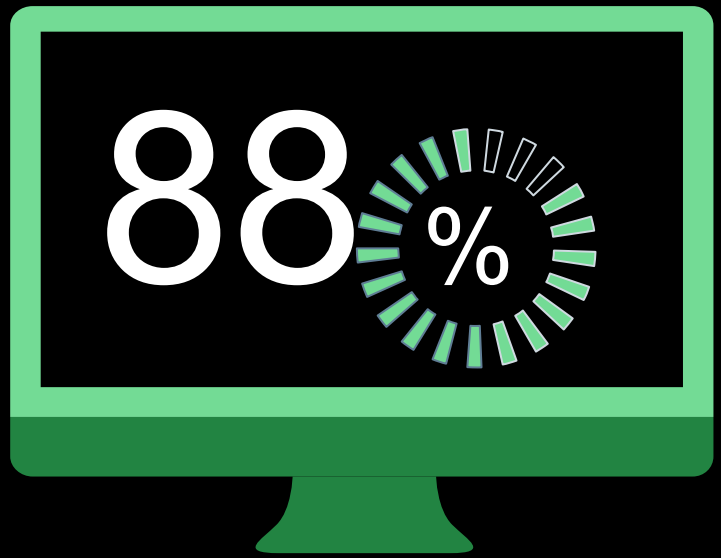
SENSITIVITY



Ratio of True Positives to All Actual Positives

Ability to correctly classify r/AMD posts

SPECIFICITY

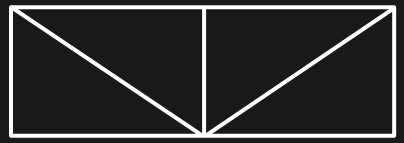


Ratio of True Negatives to All Actual Negatives

Ability to correctly classify r/NVIDIA posts

Models Summary

	TEST ACCURACY	AUC	Sensitivity | Specificity
Decision Tree	76%	87%	56% | 96%
Multinomial Naive Bayes	86%	95%	82% | 90%
Logistic Regression	**87%**	**97%**	**87% | 88%**
	Highest!	**Highest!**	**Most balanced and high overall!**



LOGISTIC REGRESSION IS THE WAY TO GO

— CHIEF HYGIENIST

04

CONCLUSION & RECOMMENDATIONS

01

02

03

04

05

06

4 STAGE PROCESS

STAGE 1 & 2 -
RECEIVE QN & CLASSIFY

To build a classifier that identifies keywords for the bot to direct customers to subreddits accurately.

STAGE 3 & 4 –
MATCH QN AND SEND LINK

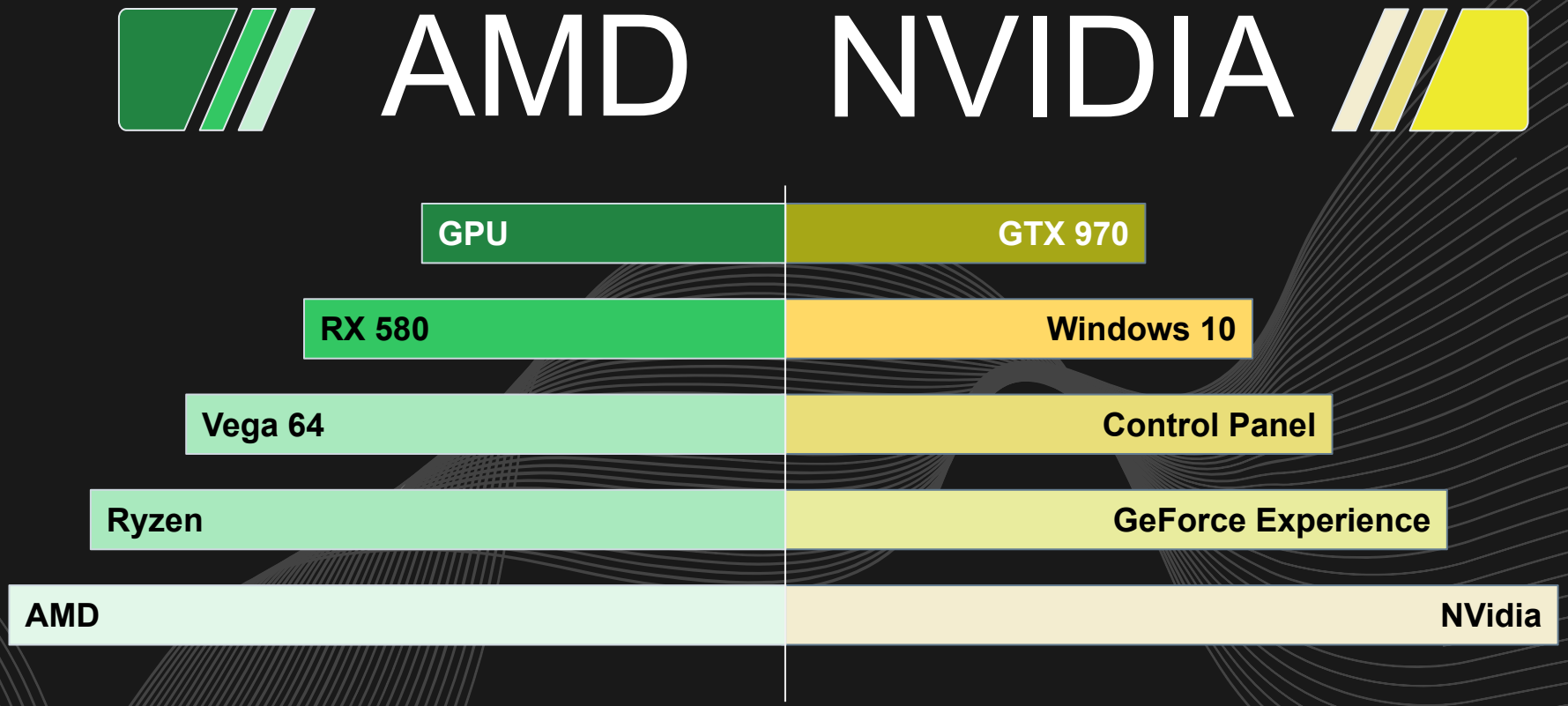
We will be identifying specific posts with solutions and assigning into a dictionary for the software team to direct customers to in the future

Good enough for Beta Test

Software engineering can proceed with using our logistic regression model

87%

TOP PHRASES FOR CLASSIFICATION



GOING FORWARD

RECOMMENDATION

Software team to get Customer ID and Product ID for automatic brand & model imputation

IMPROVE ACCURACY

Test with more models and tune more hyperparameters

ONE BOT TO RULE THEM ALL

Complete AI recommendation solution for all customer inquiries

ADDITIONAL DATA POINTS

Scrape additional solutions from other NVidia and AMD forums

THANK YOU

Any questions?