# 1800-READMES: A META RESEARCH DIVE

By Brock Green, Jeremiah Toribio, Ramiro Lopez

Oct 24, 2023

# Overview

EXECUTIVE SUMMARY

FINDINGS

RECOMMINDATIONS

Conclusion

# EXECUTIVE SUMMARY

## GOAL

Create a model that can predict the main programming language of a repository

# EXECUTIVE SUMMARY

| GOAL | BIG IDEA |
| --- | --- |
| Create a model that can predict the main programming language of a repository | Given the text of a repo's ReadMe file, one can use NLP to accurately predict a repositories programming language |

# EXECUTIVE SUMMARY

| GOAL | BIG IDEA | FINDINGS |
| --- | --- | --- |
| Create a model that can predict the main programming language of a repository | Given the text of a repo's ReadMe file, one can use NLP to accurately predict a repositories programming language | • Unigrams<br>• Bigrams<br>• Trigrams |

# EXECUTIVE SUMMARY

| GOAL | BIG IDEA | FINDINGS | RECOMMENDATIONS |
| --- | --- | --- | --- |
| Create a model that can predict the main programming language of a repository | Given the text of a repo's ReadMe file, one can use NLP to accurately predict a repositories programming language | • Unigrams<br>• Bigrams<br>• Trigrams | • Run this model to aid recruiters<br>• Compare to other departments |

**82%**

REPOS ARE PYTHON

**18%**

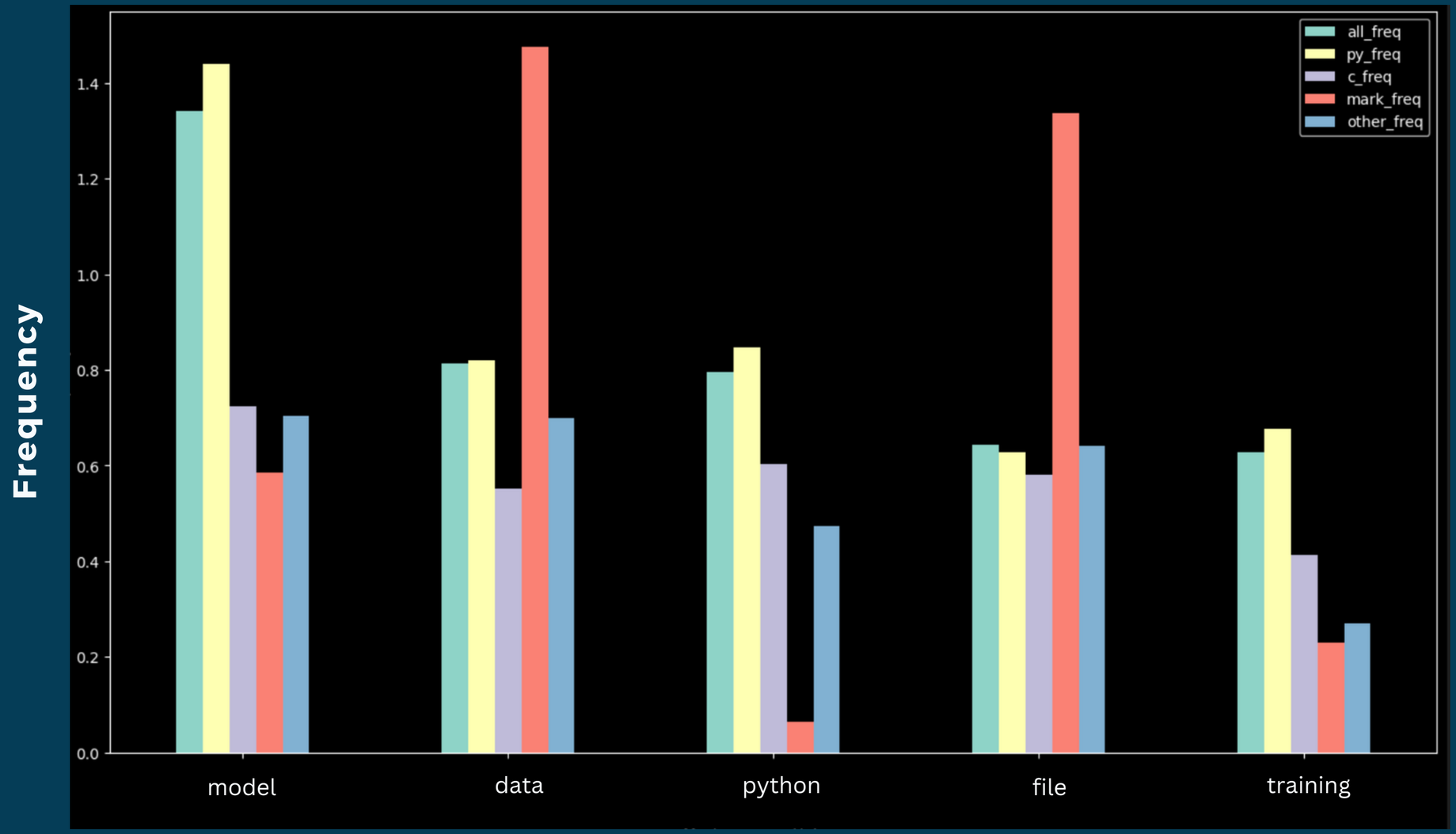
OTHER REPOS

*C++, Markdown and other*

# Unigrams

- **Markdown**
  - Dataset
  - license
  - file
  - data
- **C++**
  - Example
  - install
- **Python**
  - Model
  - python

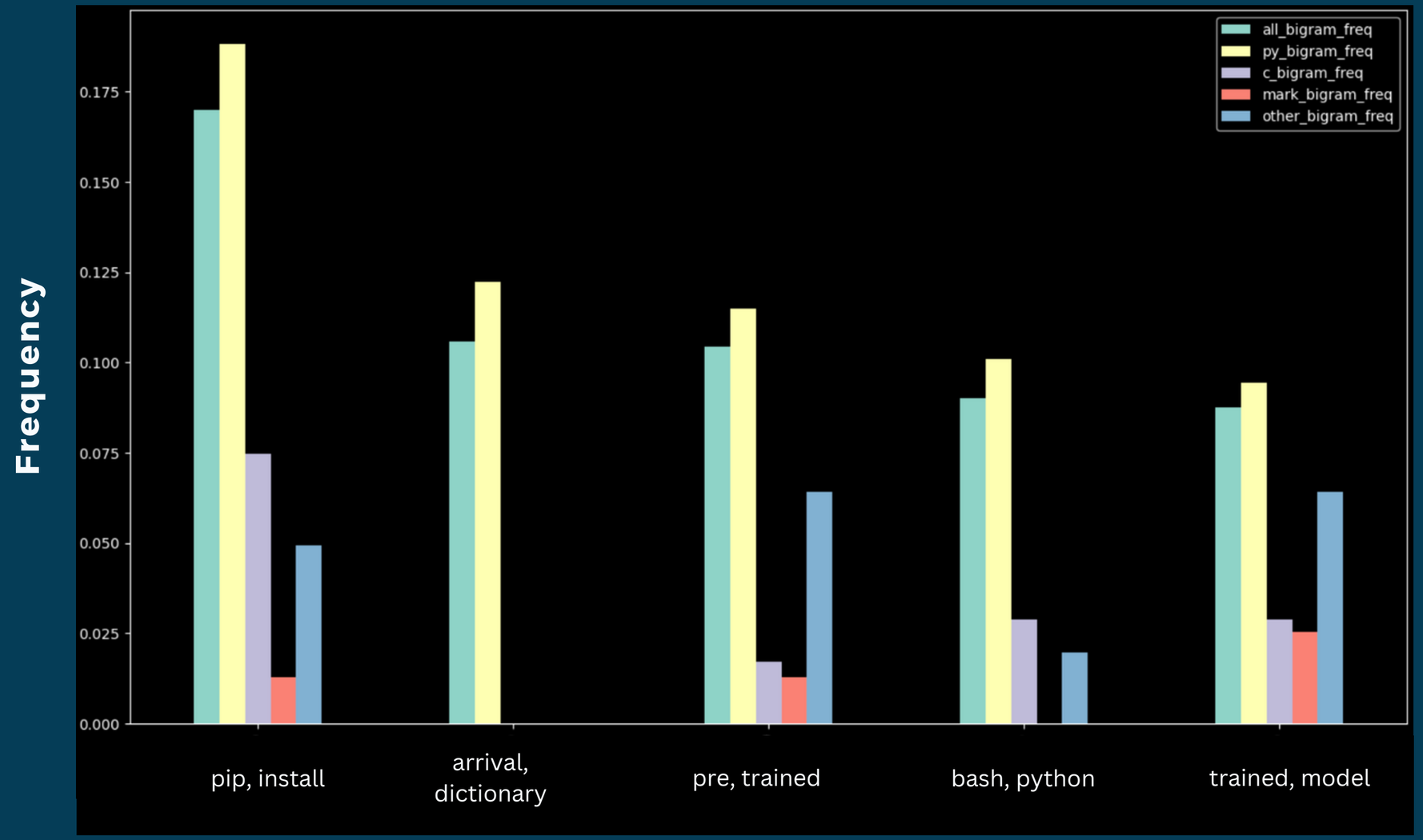
## Most common words by category



# Bigrams

- **Markdown**
  - License, File
  - Blob, Main
  - cc, nc
- **C++**
  - Git, Clone
  - Following, comand
- **Python**
  - Pip, Install
  - Pre, Trained
  - Trained, Model

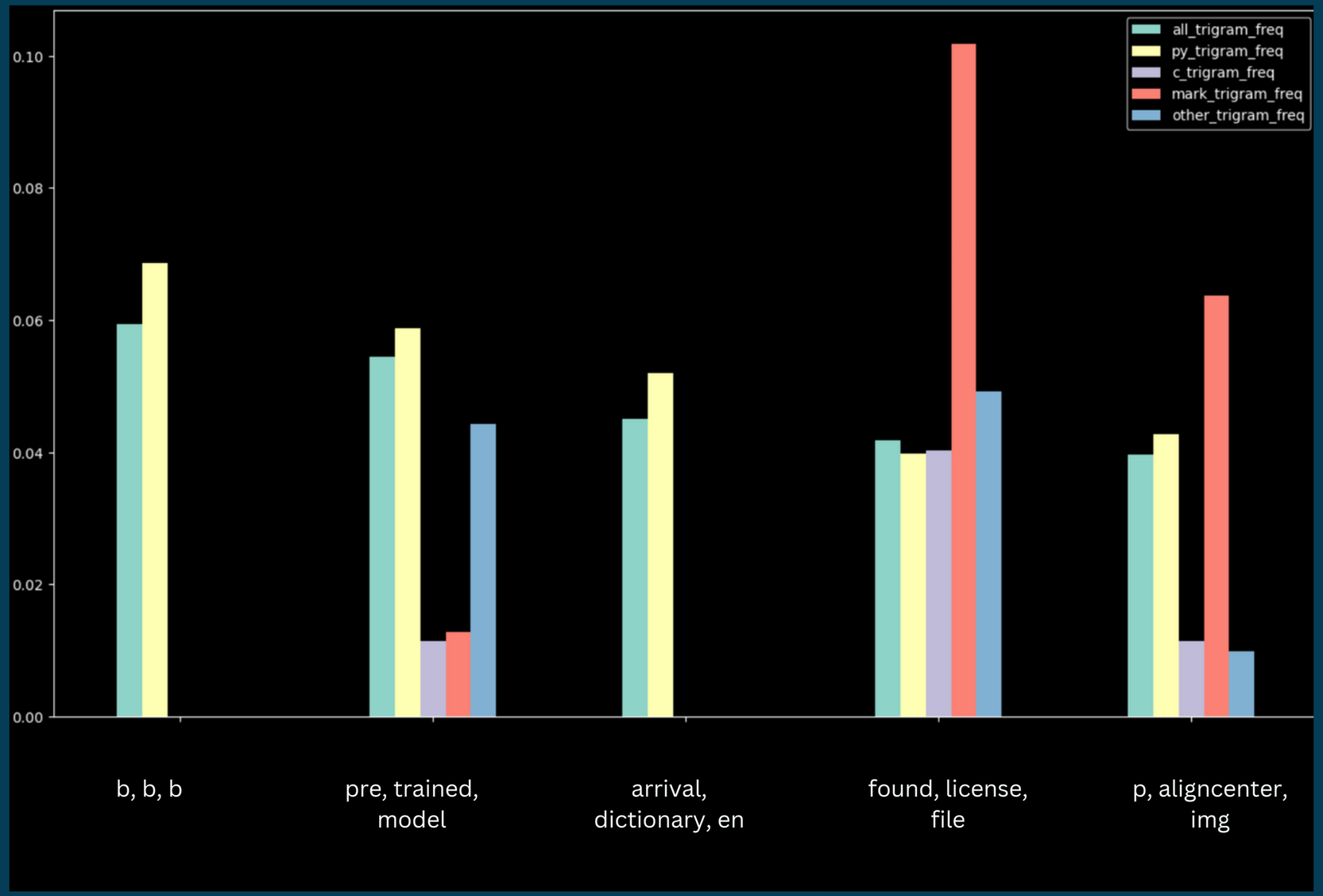
## Most common bigrams by category



# Trigrams

- **Markdown**
  - found, license, file
  - testhttps, arrival, dictionary
  - p, aligncenter, img
- **C++**
  - Git, Clone, http
- **Python**
  - pre, trained, model

## Most common trigram by category

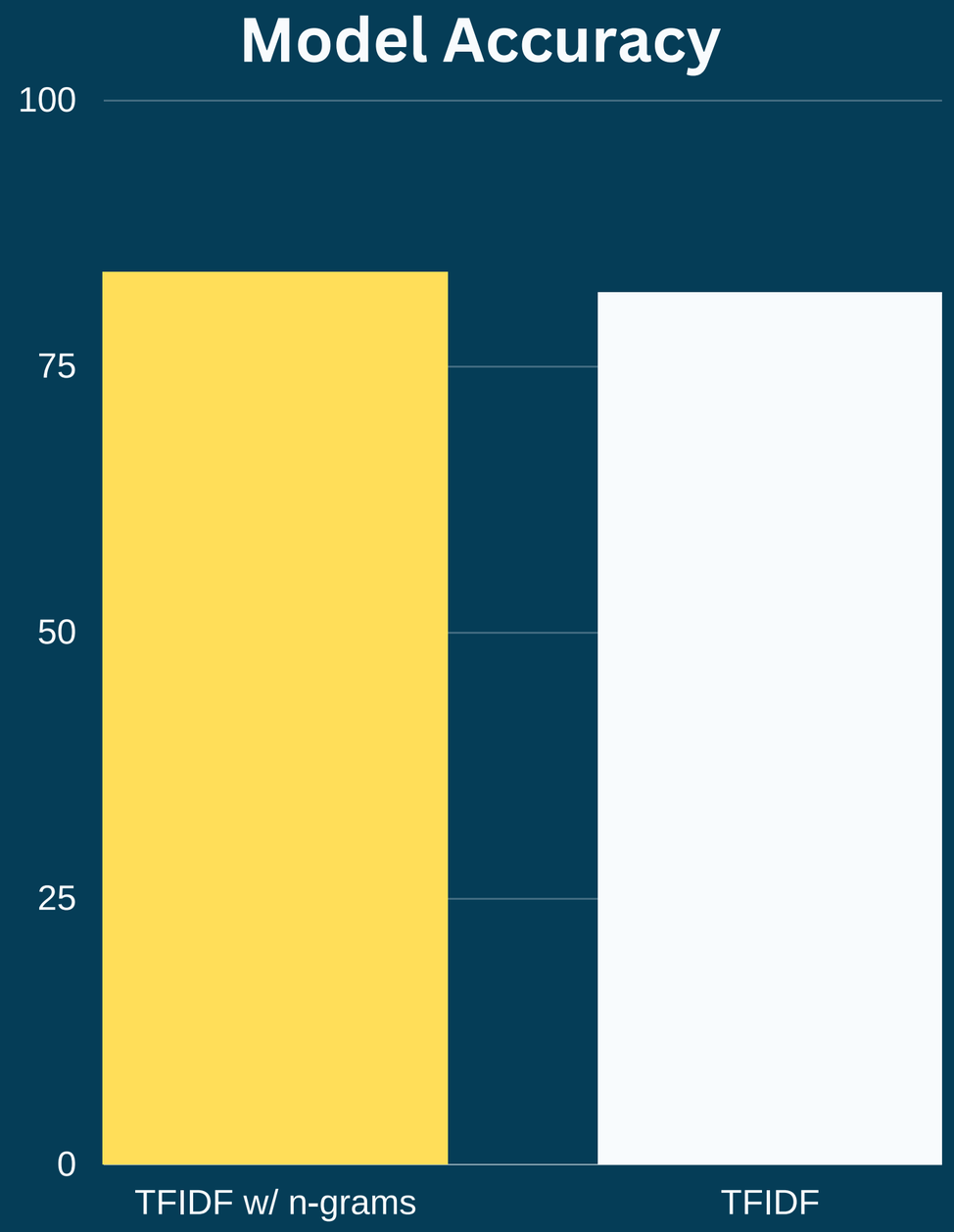


# Modeling

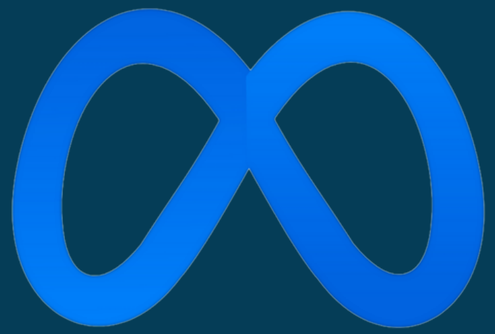
**With and without n-grams**

**Decision Tree** *of* **max_depth: 2**

Baseline - 82%
Train - 84.39%
Test - 83.75%



# Recommendations



- Assisting with team manager placements
- Potentially scan an applicants GitHub to see candidate likelyhood for Research team

# Conclusion

## MODEL ACCURACY

- +2% Better than baseline

## TAKEAWAYS

- Research teams primarily use Python
- Potential candidates can be vetted on their knowledge and experience with respective top languages

## NEXT STEPS

- Add more repositories from different departments to accurately predict for any department
- Prescriptive Model: Identify the features that would be easiest for the business to take action

# THANK YOU FOR YOUR TIME

## BROCK GREEN

Data Scientist
https://github.com/brock-green

## JEREMIAH TORIBIO

Data Scientist
https://github.com/jeremiah-toribio

## RAMIRO LOPEZ

Data Scientist
https://github.com/Ramiro-Lopez