Coding Language Predictions in Google Github README's

NLP-Classification Project

Amanda Gomez

11 MAY 2020

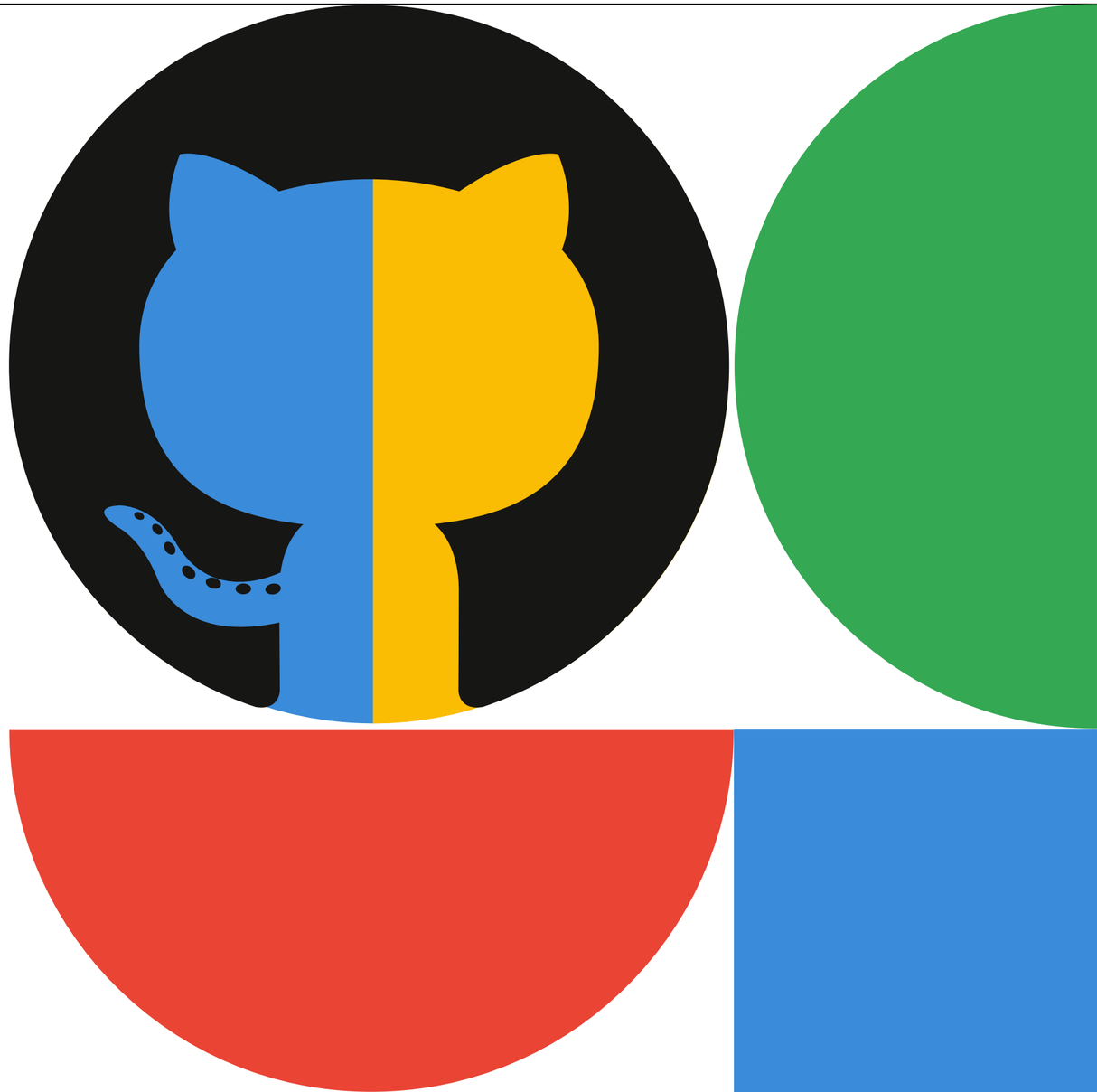
Today's Agenda

Executive Summary

Data Analysis

Conclusion

Appendix



Executive Summary

Goal

The goal for this project is to create a model that will accurately predict the primary coding language of a Github Repository given text from a README.

Data Set

This data set was scraped from Google's Github site. Not all pages were obtained due to empty repositories, although the csv used contains 1020 observations.

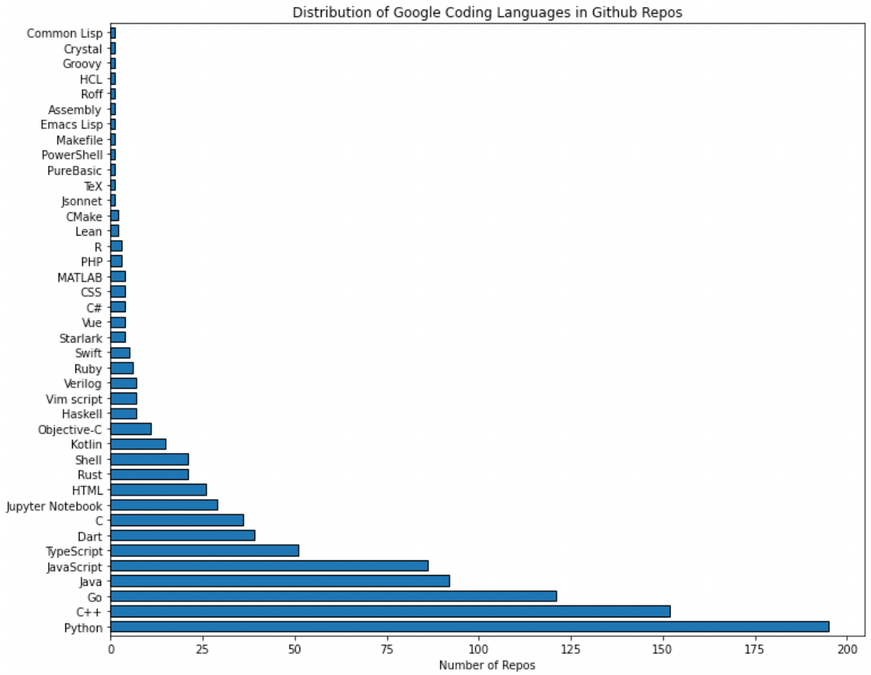
Findings

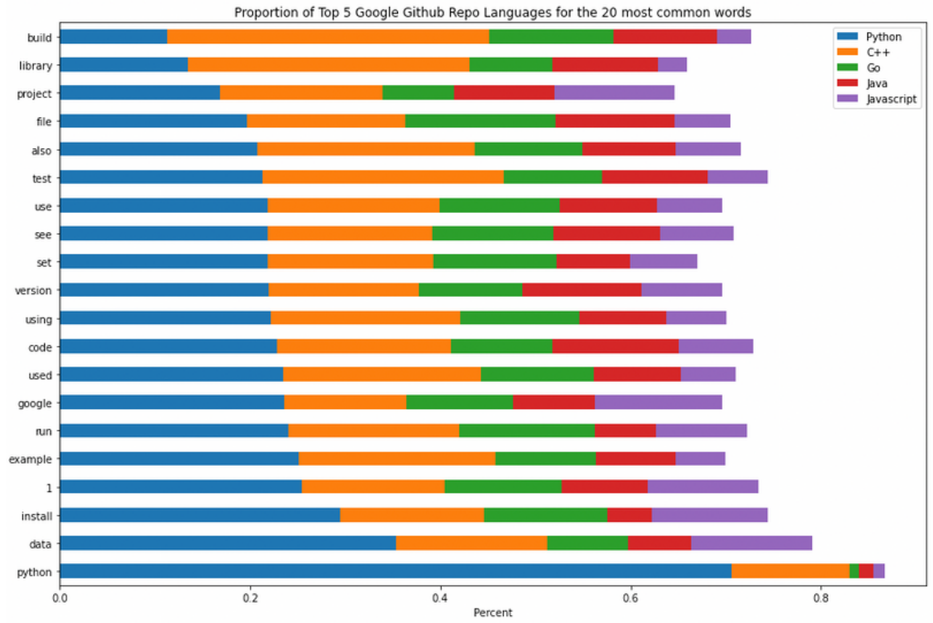
- Google uses over 30 different languages in their repos, which drove down the accuracy.
- The top common words used across the languages were verbs.

Model Performance

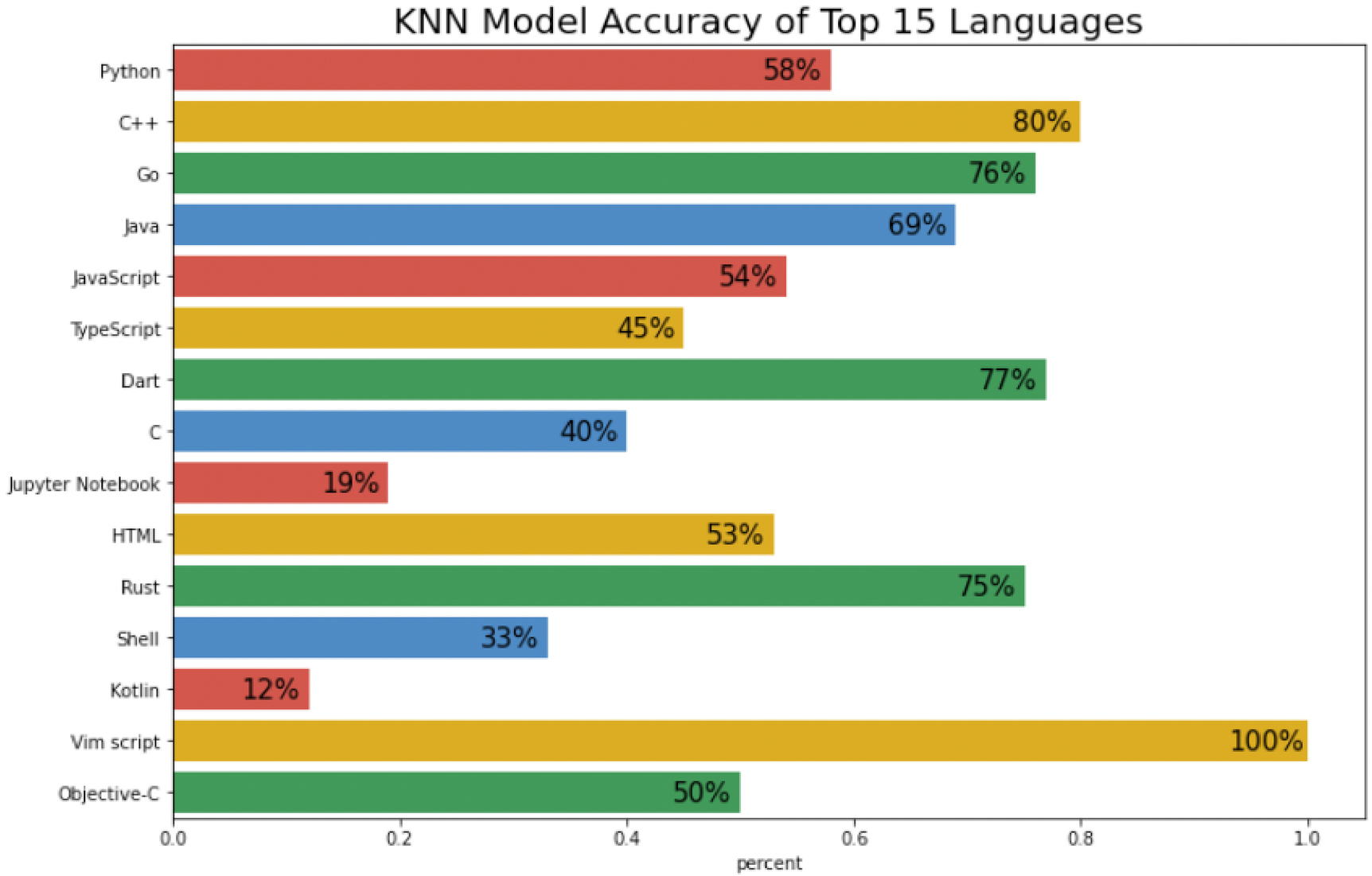
Use of my model **out performs the baseline by 28%** when predicting a coding language based on a repository's README file.

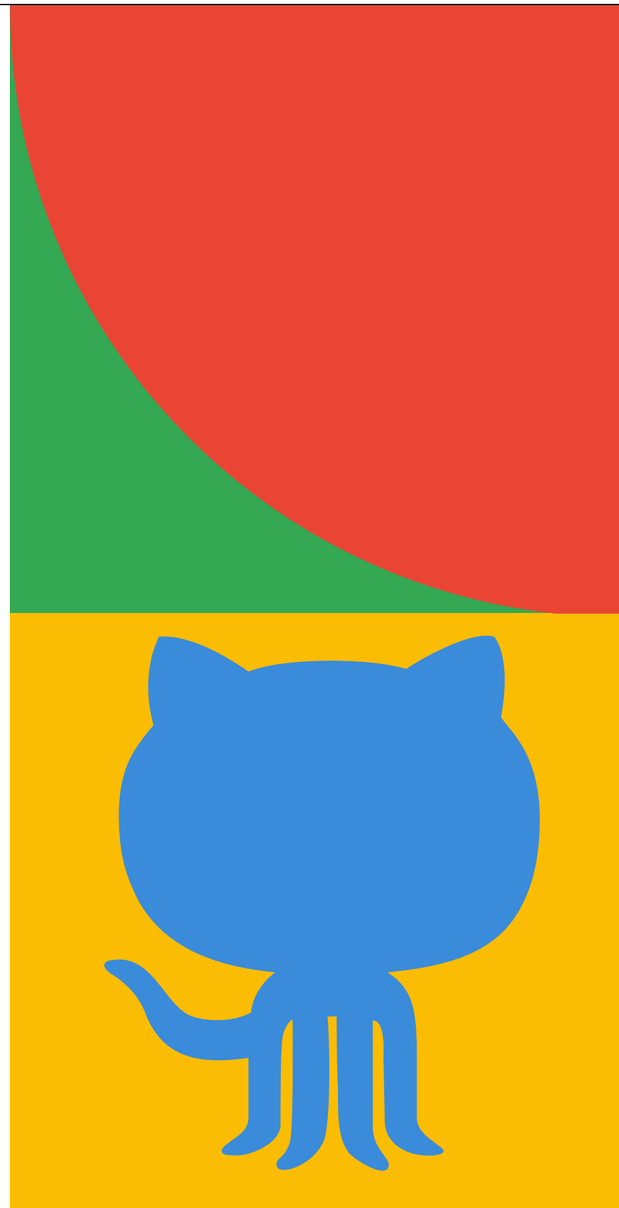
Data Analysis





Data Analysis





Conclusion

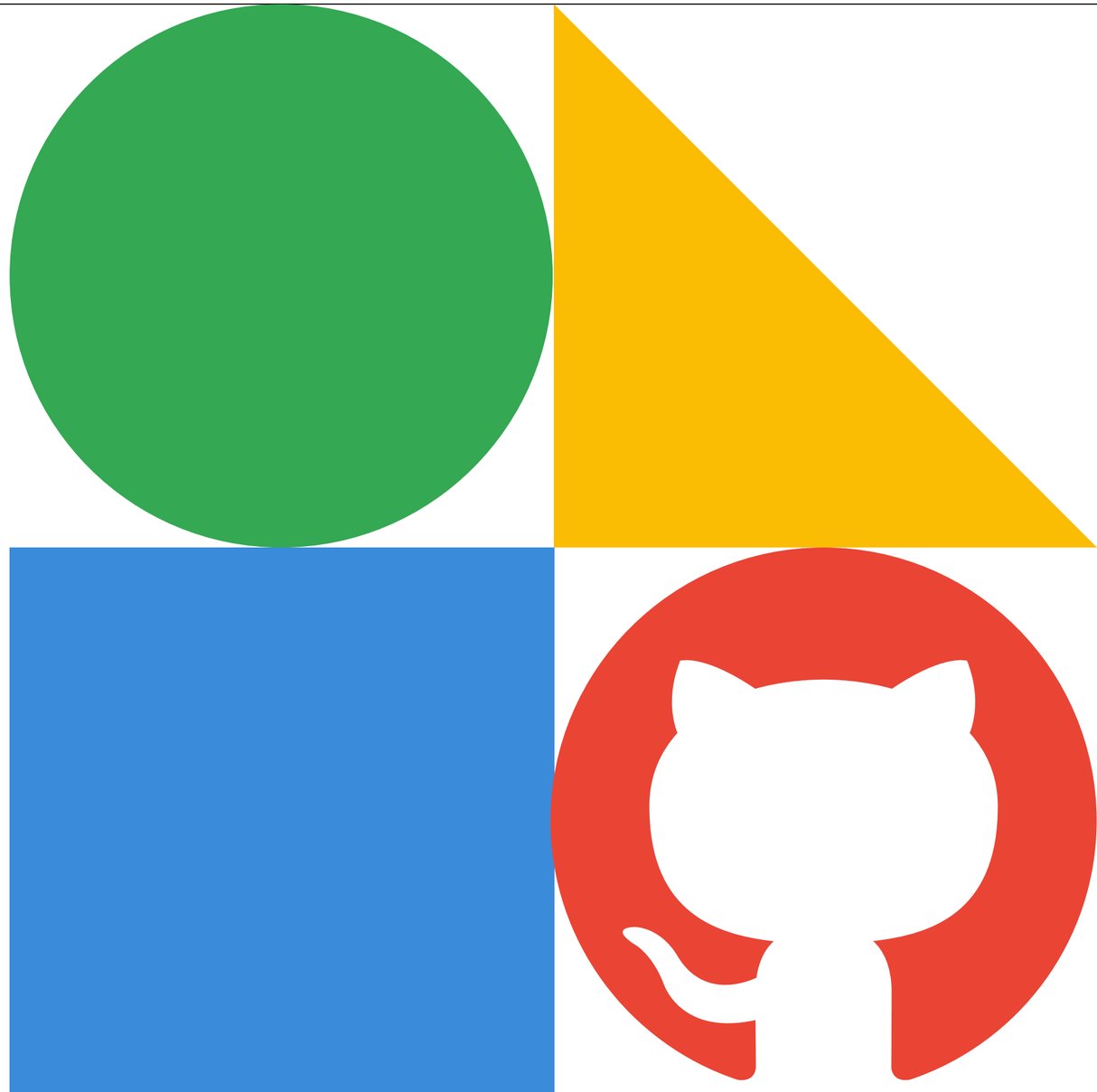
KNN predictive model out performed the baseline model by 28%

- This is likely due to the fact that Google Github repositories use 40 different programming languages.

- Most of which are verbs

Next Steps

Given more time, I'd like to filter my scraping for a limited amount of languages to identify keywords for each language.



Appendix

DATA DICTIONARY

column_name	description	key	dtype
`repo`	Link suffix in `username` / `repo_name` formatting.		object
`language`	Primary coding language used in the repo.		object
`readme_contents`	String of text scraped from repo's README file.		object
`readme_length`	Length of `README` text		int64
`clean_content`	String of `README` text that has been cleaned by `clean()` function		object
`cleaned_length`	Length of `clean_content` text		int64

GITHUB

https://github.com/o0amandagomez0o/nlp_project-readme_prediction

Thank you!

Any questions or comments?