PROGRAMMING PREDICTOR

Jacob Kulik and David Pogrebitskiy

OUR GROUP

Jacob Kulik

Data Science and Finance

Class of ‘25

David Pogrebitskiy

Data Science, Mathematics Minor

Class of ‘25

TABLE OF CONTENTS

01 Background

02 Intro to Data

03 Data Science Approach

04 Results and Next Steps

01

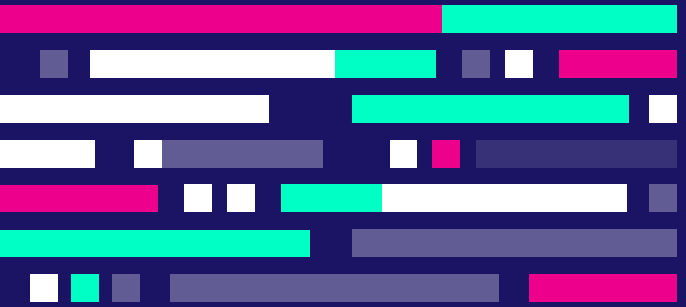
BACKGROUND

INTRODUCTION

Have you ever seen a snippet of code and wondered what language it was written in?

02

INTRO TO DATA



LABELED CODE SNIPPETS DATASET

Two column CSV file made up of the GitHub Repository dataset

FEATURES OF THE DATASET

34

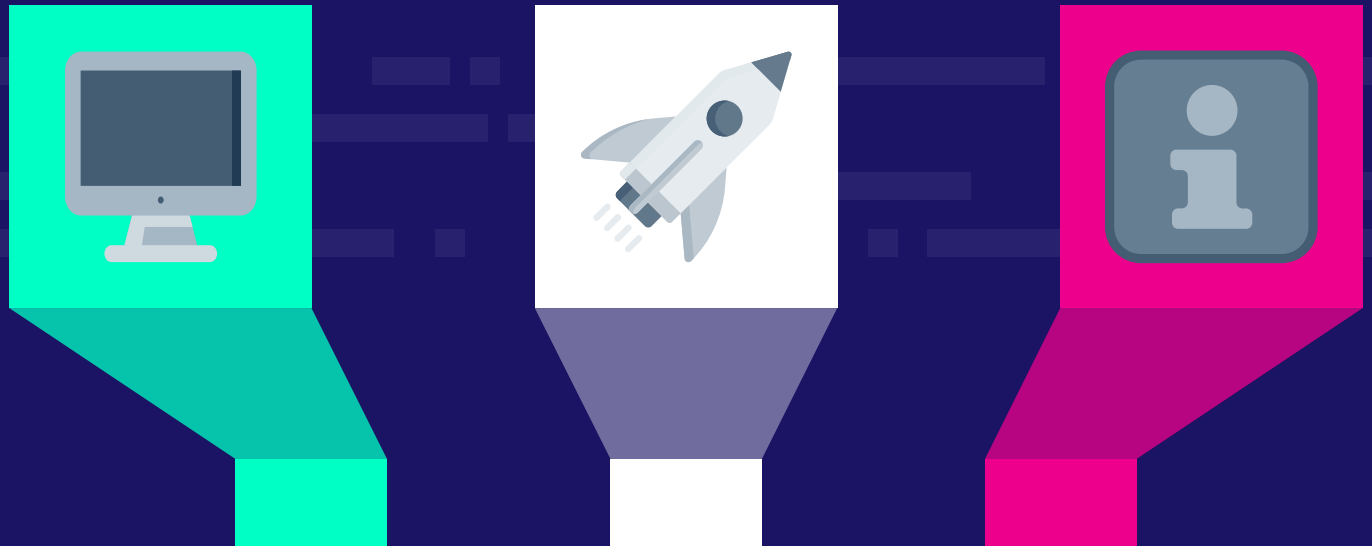
Programming Languages Included

131604

Snippets of Code

1.35

GB csv file



03

DATA SCIENCE APPROACH

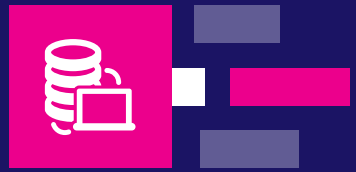
ML PIPELINE

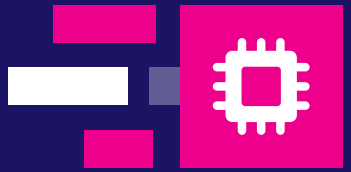
Regex

Tokenizing code while taking into account syntactical styles of programming languages

TF-IDF

Vectorizing and weighting tokens based on discriminatory characteristics



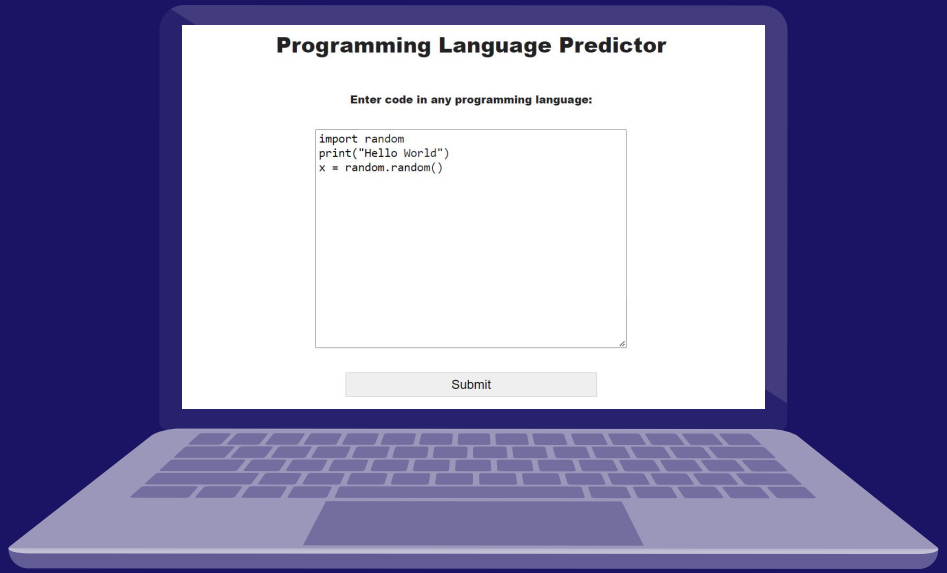


MultinomialNB

Classifying tokens based on conditional probabilities derived from Bayes' Theorem

User Input

Running **user-inputted** string through the pipeline to get most likely languages



FLASK WEB APP

User friendly UI that takes in a snippet of code and predicts the language based off our algorithm

WEB APPLICATION

Flask

Web framework allowing for a combination of Python programs and basic HTML/CSS/PHP

Pickling

Training the model once and converting it into a byte stream, which then gets unpacked in the app

04

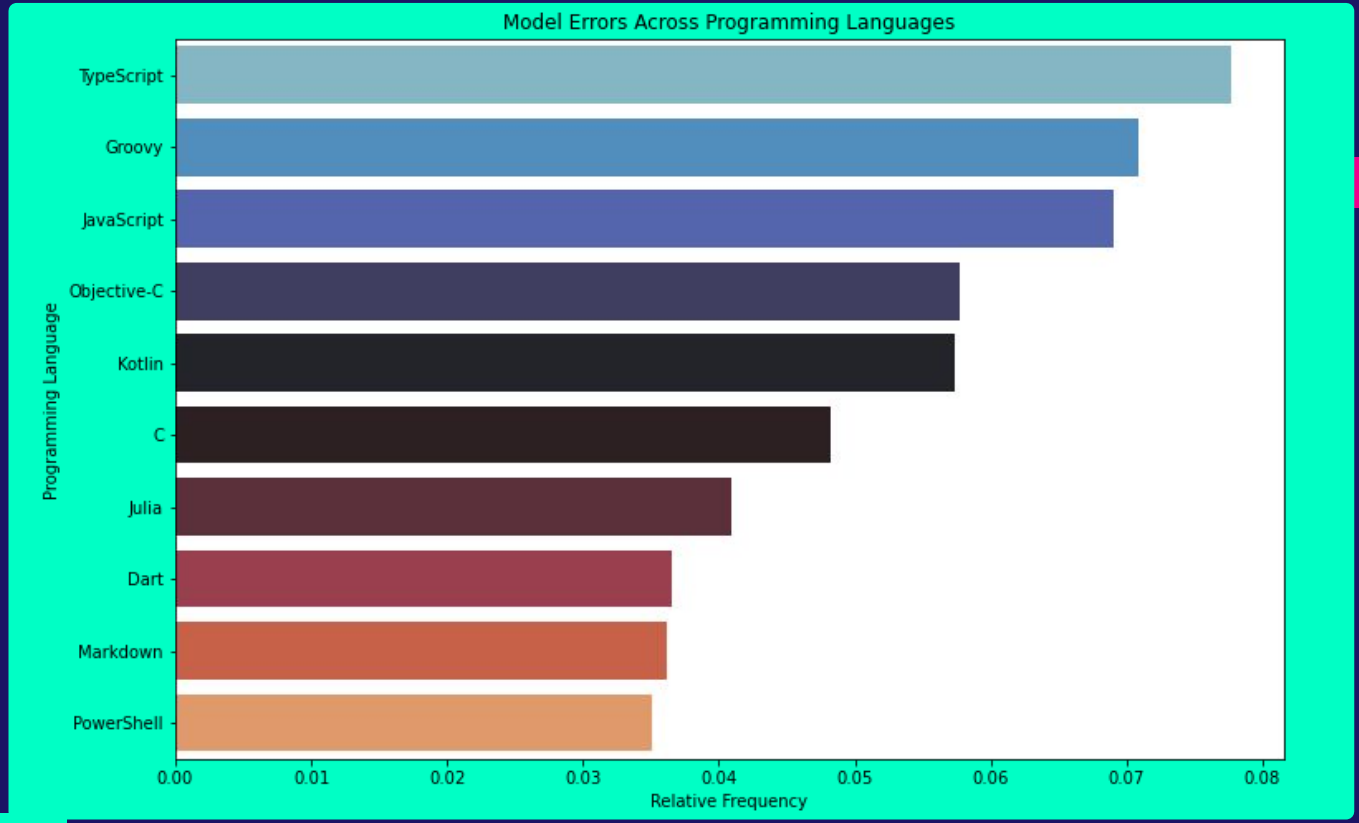
RESULTS AND CONCLUSIONS

0.90 & 0.82

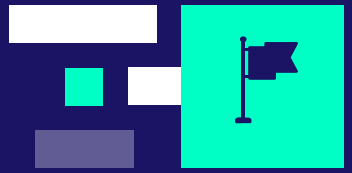
PRECISION

RECALL

MODEL ERRORS



MODEL ERROR EXPLANATION

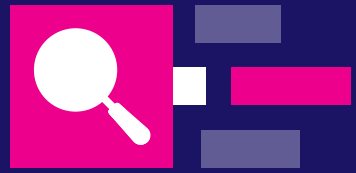


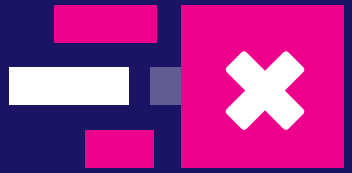
Similarities

Syntax intersections, such as semicolons at the end of every line

Human Error

Typos and accidental characters can be interpreted incorrectly



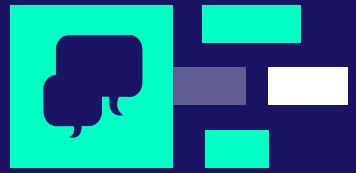


Variable Names

The algorithm cannot distinguish between variable names and built in functions

Length

Our model works best with long snippets of code because of the increase in discriminatory features

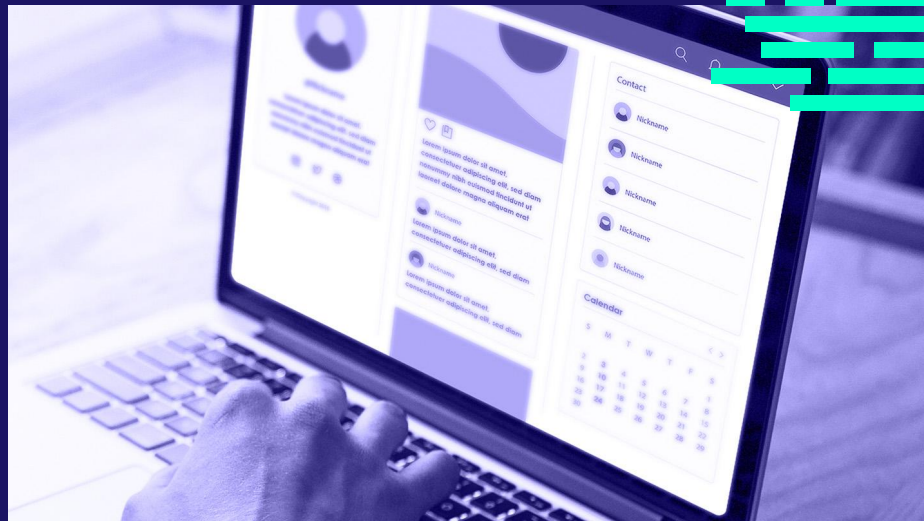


TEST IT FOR YOURSELF!

program-predictor.herokuapp.com

Future Work

- Improving UI/UX of site
- Testing accuracy-improving metric changes
- Applying text pre-processing to the data
- Accounting for user error



Thanks!

program-predictor.herokuapp.com