DSI30 Project 3

Classification Model

by Cheong Hao Ming

TABLE OF CONTENTS

01

Introduction

Background
Problem Statement

Running Lab

- Running Specialty Store
- Covid-19 increase interest in sports
- Targeting First Time Marathoners

Question?

How many KM(s) do you run in a week?

Why Marathon Runners?

Week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	Rest	4.8 km run	4.8 km run	4.8 km run	Rest	9.7	Cross
2	Rest	4.8 km run	4.8 km run	4.8 km run	Rest	11.3	Cross
3	Rest	4.8 km run	6.4 km run	4.8 km run	Rest	8.1	Cross
4	Rest	4.8 km run	6.4 km run	4.8 km run	Rest	14.5	Cross
5	Rest	4.8 km run	8.1 km run	4.8 km run	Rest	16.1	Cross
6	Rest	4.8 km run	8.1 km run	4.8 km run	Rest	11.3	Cross
7	Rest	4.8 km run	9.7 km run	4.8 km run	Rest	19.3	Cross
8	Rest	4.8 km run	9.7 km run	4.8 km run	Rest	Rest	Half Marathon
9	Rest	4.8 km run	11.3 km run	6.4 km run	Rest	16.1	Cross
10	Rest	4.8 km run	11.3 km run	6.4 km run	Rest	24.1	Cross
11	Rest	6.4 km run	12.9 km run	6.4 km run	Rest	25.7	Cross
12	Rest	6.4 km run	12.9 km run	8.1 km run	Rest	19.3	Cross
13	Rest	6.4 km run	14.5 km run	8.1 km run	Rest	29	Cross
14	Rest	8.1 km run	14.5 km run	8.1 km run	Rest	22.5	Cross
15	Rest	8.1 km run	16.1 km run	8.1 km run	Rest	32.2	Cross
1	Rest	8.1 km run	12.9 km run	6.4 km run	Rest	19.3	Cross
17	Rest	6.4 km run	9.7 km run	4.8 km run	Rest	12.9	Cross
18	Rest	4.8 km run	6.4 km run	3.2 km run	Rest	Rest	Marathon

18 Weeks Novice Marathoner Training Plan

788KM!

Question?

How many running shoes do you have?

Why Marathon Runners?

$219

/ Easy Days

For slow-paced and recovery runs

$259

/ Workout

For high tempo and HIIT style running workouts

$399

/ Raceday

High performance racing shoes

02

The Data

Collection, Cleaning, EDA

DATA COLLECTION

- Running Subreddit
- Firstmarathon Subreddit
- 18 iterations per subreddit (pushshift.io limit at 100)
- 1800 submission per subreddit

DATA CLEANING

/ Null Values

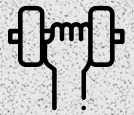
Selftext + Title
Drop Null Values
Drop Duplicates

/ Lemma

Lemmatize words

/ Remove

URLs
HTML
Non-Letters

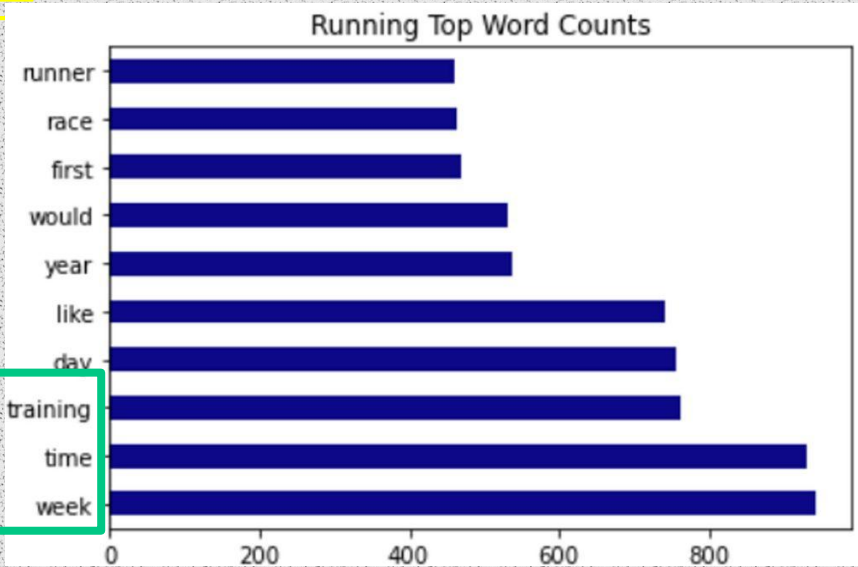


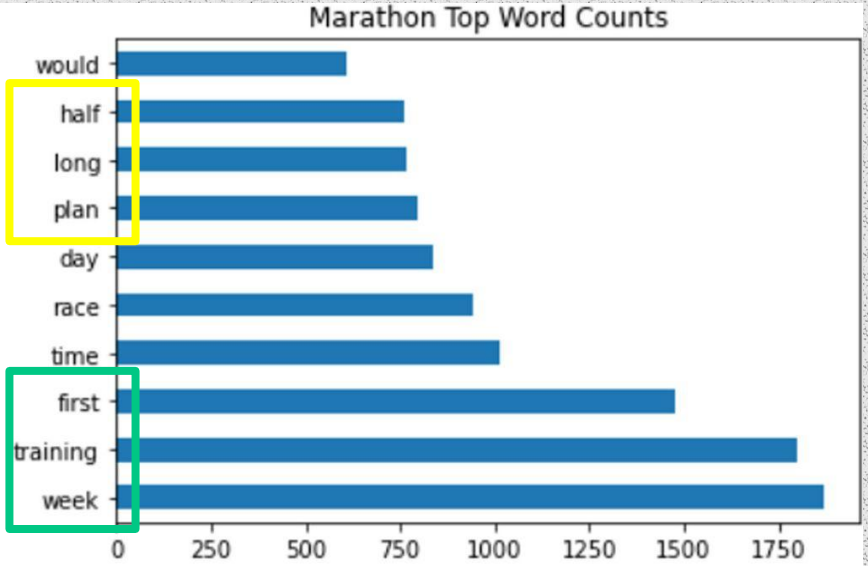
/ Stopwords

Additional stopwords from EDA

```
["run", "running", "marathon",
 "get", "wa", "mile"]
```

EDA


Running Top Word Counts
runner
race
first
would
year
like
day
training
time
week
0
200
400
600
800


Marathon Top Word Counts
would
half
long
plan
day
race
time
first
training
week
0
250
500
750
1000
1250
1500
1750

03

Modelling

Modelling Process
Evaluation

MODELS


/ CountVectorizer


GYM
/ TfidfVectorizer


/ Naive Bayers


/ K Nearest Neighb.

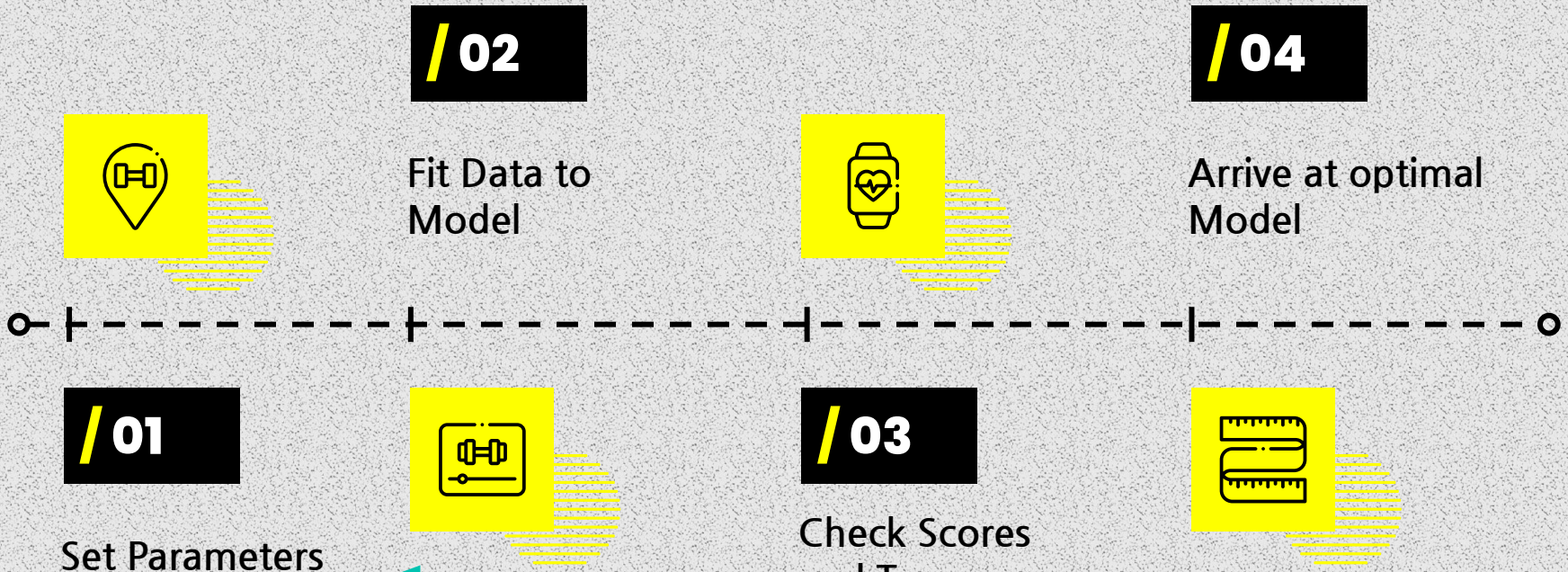

/ Logistic Reg.


/ Bagging


/ Random Forest


10
Models
(Including baseline model)

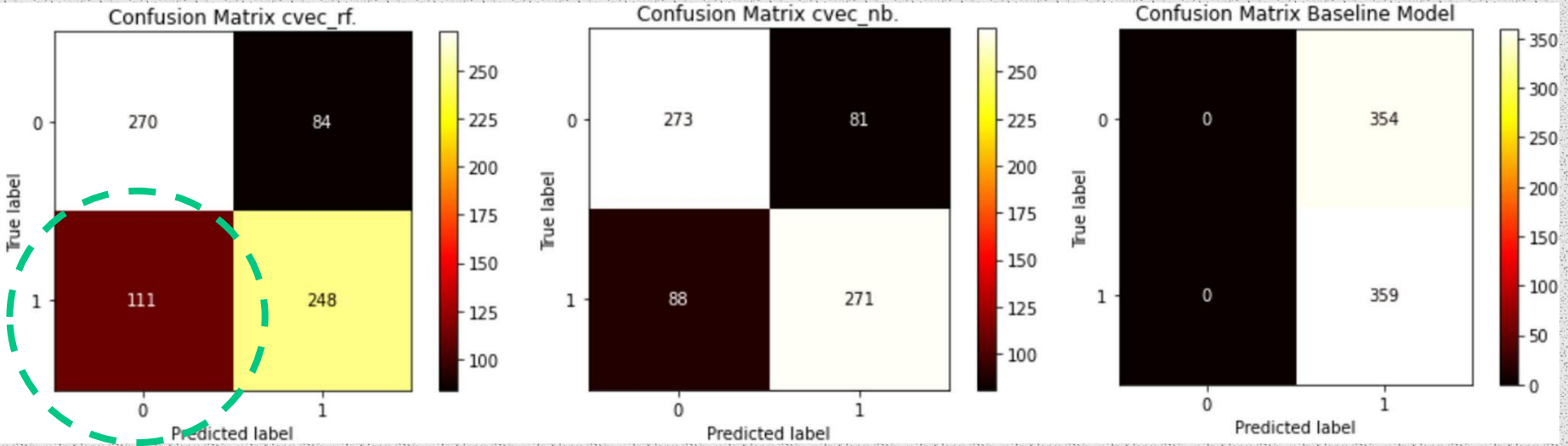
MODELLING PROCESS



MODEL EVALUATION

	Vectorizer	Classifier	Train	Test	Train-Test	Best Score	Specificity	Sensitivity
1	Baseline	Baseline	NA	NA	NA	0.503	0.0	1.0
2	Count Vectorizer	Naive Bayers	0.859	0.763	0.096	**0.802**	0.771	0.755
3	Tfidf Vectorizer	Naive Bayers	0.879	0.762	0.117	0.795	0.749	0.774
4	Count Vectorizer	Logistic Regression	0.920	0.767	0.153	0.789	0.740	0.794
5	Tfidf Vectorizer	Logistic Regression	0.895	0.774	0.121	0.789	**0.799**	0.749
6	Count Vectorizer	K Nearest Neighbor	0.746	0.620	0.126	0.627	0.342	**0.894**
7	Tfidf Vectorizer	K Nearest Neighbor	0.676	0.539	0.137	0.545	0.257	0.816
8	Count Vectorizer	Bagging	0.980	0.718	0.262	0.732	0.698	0.738
9	Count Vectorizer	Random Forest	0.791	0.727	**0.064**	0.754	0.777	0.652
10	Tfidf Vectorizer	Random Forest	0.797	0.715	0.082	0.754	0.790	0.643

MODEL EVALUATION



04

Conclusion

Recommendation
Future Works

Naive Bayer (CountVectorizer) Classifier Model

1. 80.2% Accuracy > 50.3% of baseline
2. No severe overfitting
3. Balanced FN and FP

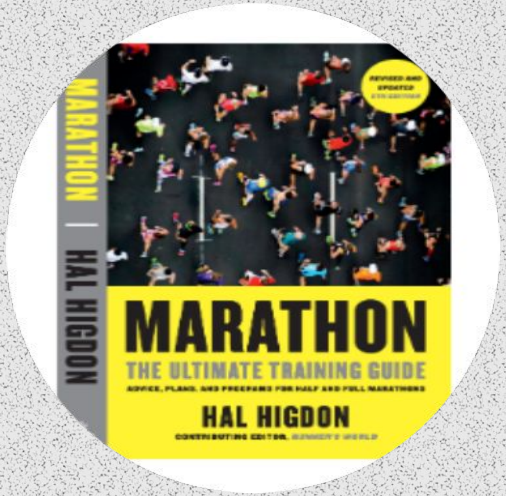
Recommendation

Naive Bayers (CVEC)		Naive Bayers (TVEC)		Random Forest (CVEC)		Random Forest (TVEC)	
	coef		coef		coef_rf		coef_rf
higdon novice	4.324131	higdon novice	2.176424	first	0.043279	first	0.050307
marathoner	3.492397	higdon	2.068963	plan	0.042406	training	0.035754
novice plan	3.464998	chicago	2.037032	higdon	0.035971	higdon	0.029987
hartford	3.282677	hal	2.002593	training	0.033548	finish	0.025818
chicago	3.100355	hal higdon novice	1.992613	hal	0.023150	week	0.024066
registered	3.016974	hal higdon	1.961368	week	0.022039	full	0.023296
following hal	2.877212	taper	1.951290	finish	0.021645	plan	0.021582
first sunday	2.825918	novice	1.812514	full	0.020133	hal higdon	0.020960
san	2.825918	first full	1.760943	recently	0.017774	hal	0.016131
parent	2.825918	miler	1.653291	finished	0.017477	higdon novice	0.016085

Recommendation

Mr. Hal Higdon
King of Running Plans

- Become stockist for his book

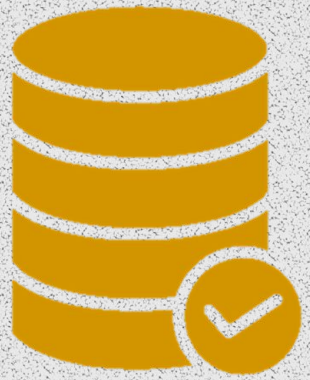


- Co-branding or collaborative opportunities

- Video-series on following his training plans

Further Works

Gather local data for more relevance



Collect more data to improve accuracy

THANKS!

“It is not the distance you must conquer in running, it is yourself!”

-Micheal D’Aulerio