The Data Science Process

Antony Ross

Silicon Valley Code Camp
October 8, 2017

The Data Science Process

1.) Identify a useful question

2.) Acquire the data

3.) Clean the data

4.) Explore the data

5.) Model the data

6.) Communicate the results

Identify a Useful Question


ONLINE BOOK
RENT
READ
RETURN
RENTAL CLUB

Customer Data

age	gender	monthly rate	membership months	cancelled
23	female	10.99	3	no
57	female	12.99	27	no
25	male	12.99	11	yes
37	male	12.99	29	no
44	male	7.99	18	no
62	female	10.99	16	yes
24	female	12.99	5	yes


COSTCO
WHOLESALE
TIRE SALES
TIRE INSTALLATION

Three Types of Returners

- Always Returners
- Never Returners
- Contingent Returners

Some of the reasons why Contingent Returners abandon their carts

- The receptacle is too far from where they've parked their car.
- They have a child whom they do not want to leave unattended.
- The weather is bad.
- They have a disability or difficulty with movement.
- Other carts are abandoned near by.

Get the Data



Clean the Data


40 YARD DASH
RB
NFL
SCOUTING
COMBINE
HYUNDAI

NFL Combines

Sample Data

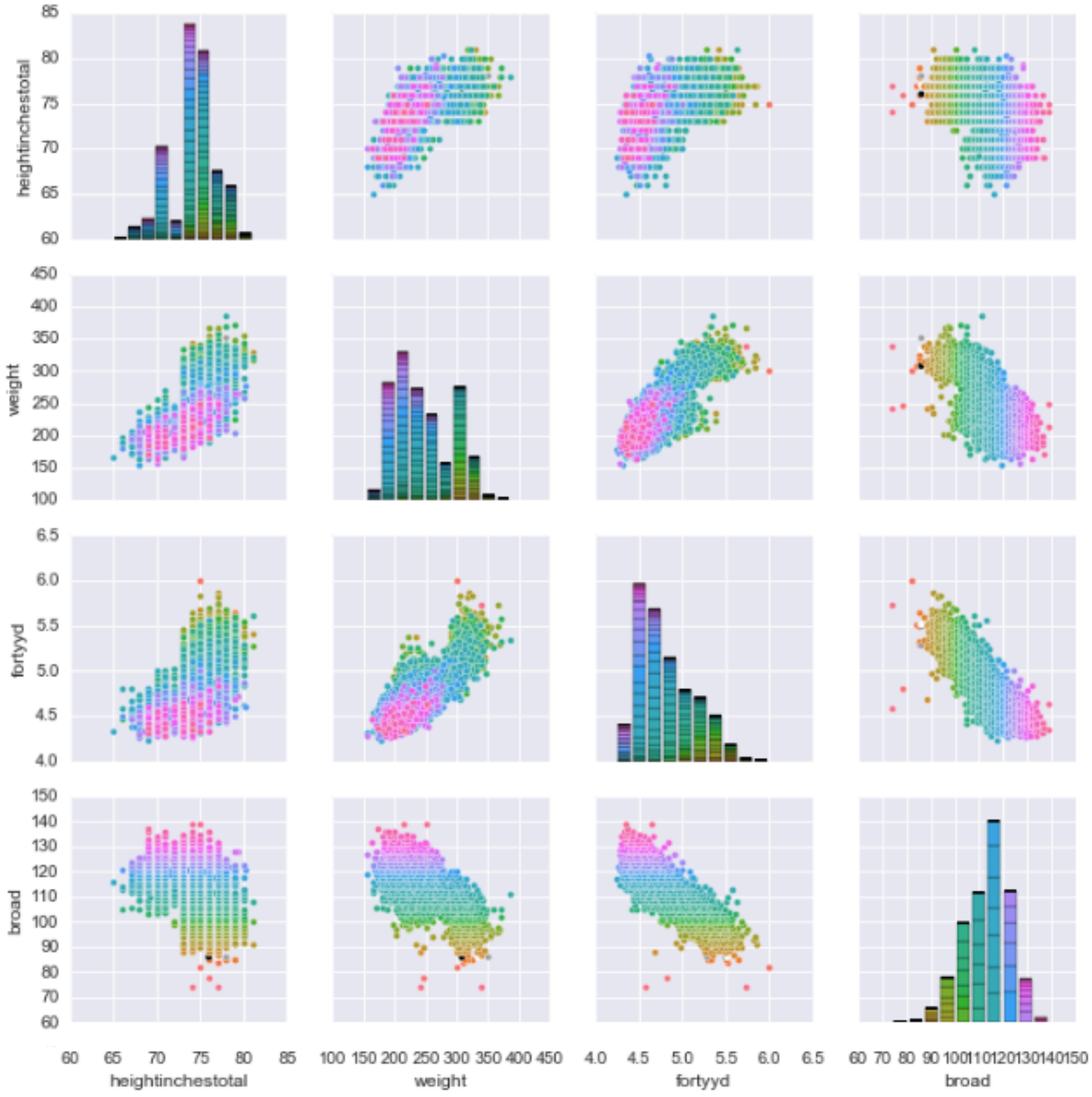
	year	name	position	heightinchestotal	weight	fortyyd	vertical	broad	bench	round	college
520	2014	Tre Mason	RB	68.0	207	4.50	38.5	126	0	3	Auburn
521	2014	Jeff Mathews	QB	76.0	223	5.26	25.5	105	0	0	Cornell
522	2014	Jake Matthews	OT	77.0	308	5.07	30.5	105	24	1	Texas A&M
523	2014	Jordan Matthews	WR	75.0	212	4.46	35.5	120	21	2	Vanderbilt
524	2014	Josh Mauro	DE	78.0	271	5.21	32.0	116	21	0	Stanford
525	2014	AJ McCarron	QB	75.0	220	4.94	28.0	99	0	6	Alabama
526	2014	Daniel McCullers	DT	79.0	352	0.00	20.5	97	27	7	Tennessee
527	2014	Dexter McDougle	CB	70.0	196	0.00	0.0	0	0	3	Maryland
528	2014	Keith McGill	CB	75.0	211	4.51	39.0	129	0	4	Utah
529	2014	Jerick McKinnon	RB	69.0	209	4.41	40.5	132	32	3	Georgia Southern

NFL Combines

Summary Data

	heightinchestotal	weight	fortyyd	threecone	vertical	broad	bench
count	4947.000000	4947.000000	4947.000000	4947.000000	4947.000000	4947.000000	4947.000000
mean	74.035476	245.579745	4.610386	1.503002	28.741257	95.944006	15.723873
std	2.614778	45.639366	0.974087	2.929683	11.596749	41.826340	10.840896
min	65.000000	155.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	73.000000	208.000000	4.530000	0.000000	28.000000	101.000000	0.000000
50%	74.000000	237.000000	4.690000	0.000000	32.500000	112.000000	18.000000
75%	76.000000	289.000000	4.990000	0.000000	35.500000	119.000000	24.000000
max	82.000000	386.000000	6.050000	8.310000	46.000000	147.000000	51.000000

Pair Plot


heightinchestotal
weight
fortyyd
broad
85
80
75
70
65
60
450
400
350
300
250
200
150
100
6.5
6.0
5.5
5.0
4.5
4.0
150
140
130
120
110
100
90
80
70
60
60 65 70 75 80 85
100 150 200 250 300 350 400 450
4.0 4.5 5.0 5.5 6.0 6.5
60 70 80 90 100 110 120 130 140 150

Clean and Transform Data

NFL Combine

height	weight	40 yd dash	year of eligibility	round drafted
68	197	4.05	Senior	3
76	253	5.02	Junior	0
77	308	0	Senior	5
75	212	4.46	Senior	3
78	271	5.21	Sophomore	2
75	242	4.94	Junior	0
76	352	0	Senior	2

NFL Combine

height	weight	40 yd dash	year of eligibility	round drafted
68	197	4.05	Senior	3
76	253	5.02	Junior	0
77	308	NaN	Senior	5
75	212	4.46	Senior	3
78	271	5.21	Sophomore	2
75	242	4.94	Junior	0
76	352	NaN	Senior	2

One-Hot Encoding

NFL Combine

height	weight	40 yd dash	year of eligibility	round drafted
68	197	4.05	Senior	3
76	253	5.02	Junior	0
77	308	NaN	Senior	5
75	212	4.46	Senior	3
78	271	5.21	Sophomore	2
75	242	4.94	Junior	0
76	352	NaN	Senior	2

NFL Combine

height	weight	40 yd dash	soph	junior	senior	round drafted
68	197	4.05	0	0	1	3
76	253	5.02	0	1	0	0
77	308	NaN	0	0	1	5
75	212	4.46	0	0	1	3
78	271	5.21	1	0	0	2
75	242	4.94	0	1	0	0
76	352	NaN	0	0	1	2

NFL Combine

height	weight	40 yd dash	soph	junior	senior	drafted
68	197	4.05	0	0	1	1
76	253	5.02	0	1	0	0
77	308	NaN	0	0	1	1
75	212	4.46	0	0	1	1
78	271	5.21	1	0	0	1
75	242	4.94	0	1	0	0
76	352	NaN	0	0	1	1

Feature Engineering

40-yard dash	Weight	Height	Drafted
5.10	290	74	1
4.92	275	75.5	1
4.43	178	69	0
4.62	221	74.5	1
4.91	248	75	0
5.53	303	77	0
4.47	189	71	1
4.56	205	71	1
4.75	267	73	0
4.84	261	74	1

Feature Engineering

40-yard dash	BMI (wt/ht^2)	Drafted
5.10	37.2	1
4.92	33.9	1
4.43	26.3	0
4.62	28	1
4.91	31	0
5.53	35.9	0
4.47	26.4	1
4.56	28.6	1
4.75	35.2	0
4.84	33.5	1


44
OL

Feature Engineering

40-yard dash	BMI (wt/ht^2)	Drafted
5.10	37.2	1
4.92	33.9	1
4.43	26.3	0
4.62	28	1
4.91	31	0
5.53	35.9	0
4.47	26.4	1
4.56	28.6	1
4.75	35.2	0
4.84	33.5	1

Feature Engineering

Speed-to-Size (40-yd/bsa)	BMI (wt/ht^2)	Drafted
2.16	37.2	1
2.06	33.9	1
2.02	26.3	0
1.97	28	1
2.23	31	0
2.00	35.9	0
2.03	26.4	1
1.99	28.6	1
1.85	35.2	0
2.03	33.5	1

Explore the Data

PLAYLIST

Discover Weekly

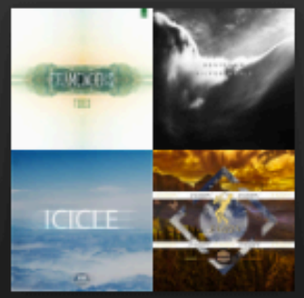
Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your favourites!

Created by: **Spotify** • 30 songs, 2 hr 48 min

PLAY FOLLOWING ••• FOLLOWER 1

Filter Download

	SONG	ARTIST	ALBUM		
+	The Sky out of Your Window	Melorman	Waves	4 days ago	3:21
+	You Have Love	Axel Thesleff	You Have Love	4 days ago	7:21
+	You're Still In It	Chihei Hatakeyama	You're Still In It	4 days ago	18:26
+	Morning Mountain	Essay	Morning Mountain	4 days ago	6:42



PLAYLIST

LIKED

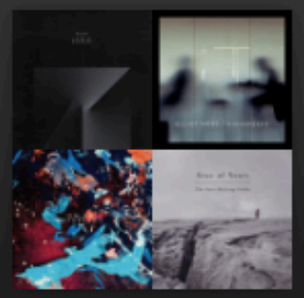
Created by: **hisbiz** • 334 songs, 24 hr 17 min

PLAY

Filter

Download

	TITLE	ARTIST
+	Breathing Light	Frameworks
+	Superior	Silver Maple
+	Icicle	AK
+	Jazzin	Flap Jack
+	Dusk	filous
+	The Way U Do	Shlohmo
+	Mirror Maru	Cashmere ...
+	Never Too Far	Sorrow
+	Mixed Signals - Synkro Remix	Frederic R...
+	Swarm	Boogrov, ...



PLAYLIST

REJECTED

Created by: **hisbiz** • 331 songs, 28 hr 44 min

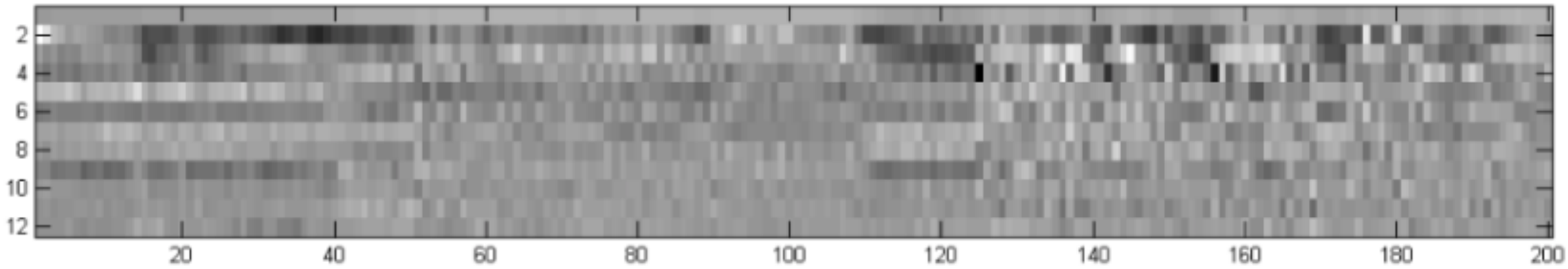
PLAY

Filter

Download

	TITLE	ARTIST
+	Frogs	Charles M...
+	Best Light	Elliot Moss
+	The Silence	Om Unit fe...
+	Passing Skies	Seas of Ye...
+	Urban Transition	Jimmy Wa...
+	Springflower	NkisOk
+	Where Did The Children Go	Deformer
+	Tactical Nuclear Penguin	Spenghead
+	Prometheus	Pythius
+	Inadequante	HE3Dless

Timbre in a selected track



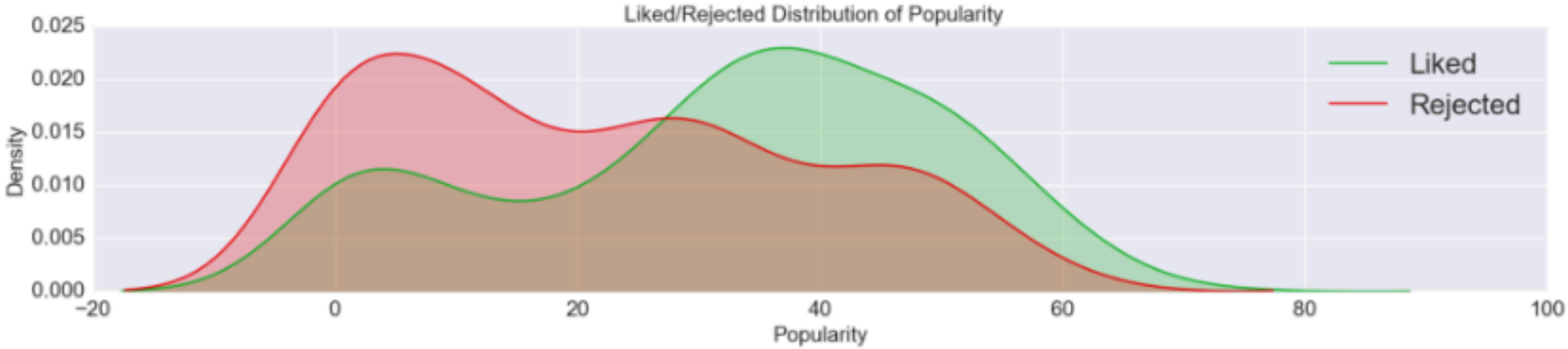
Select Features

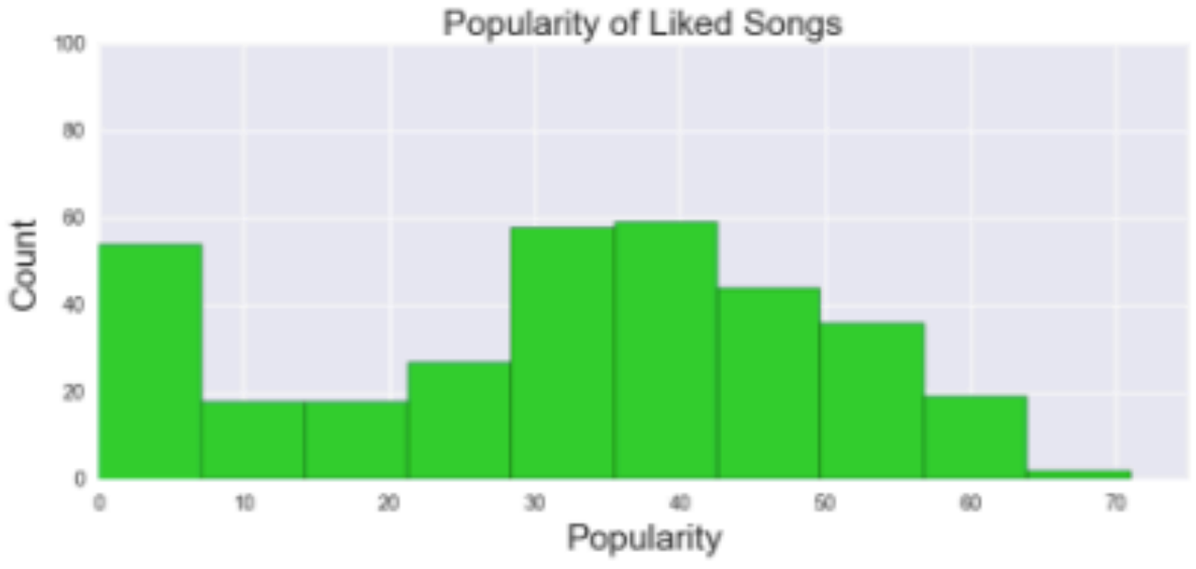
Song	Duration	Pitch	Timbre	Tempo	Popularity	Genre

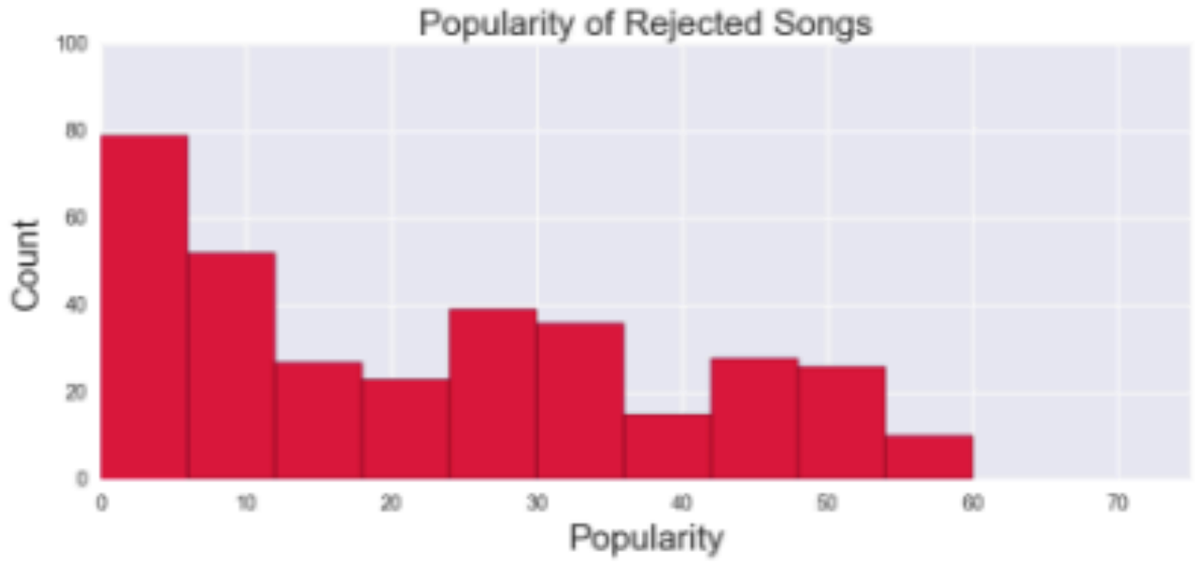
Select Features

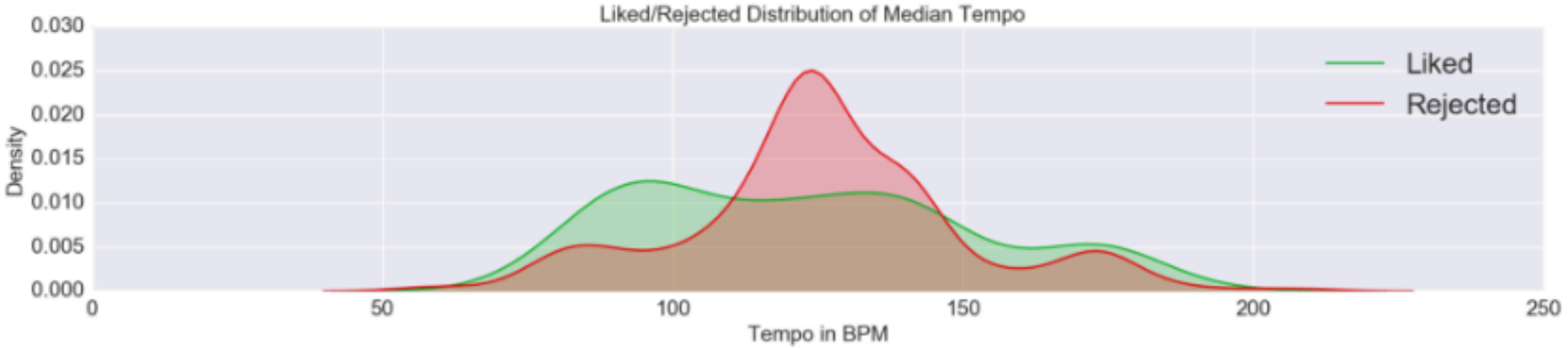
Song	Duration	Pitch	Timbre	Tempo	Popularity	Genre

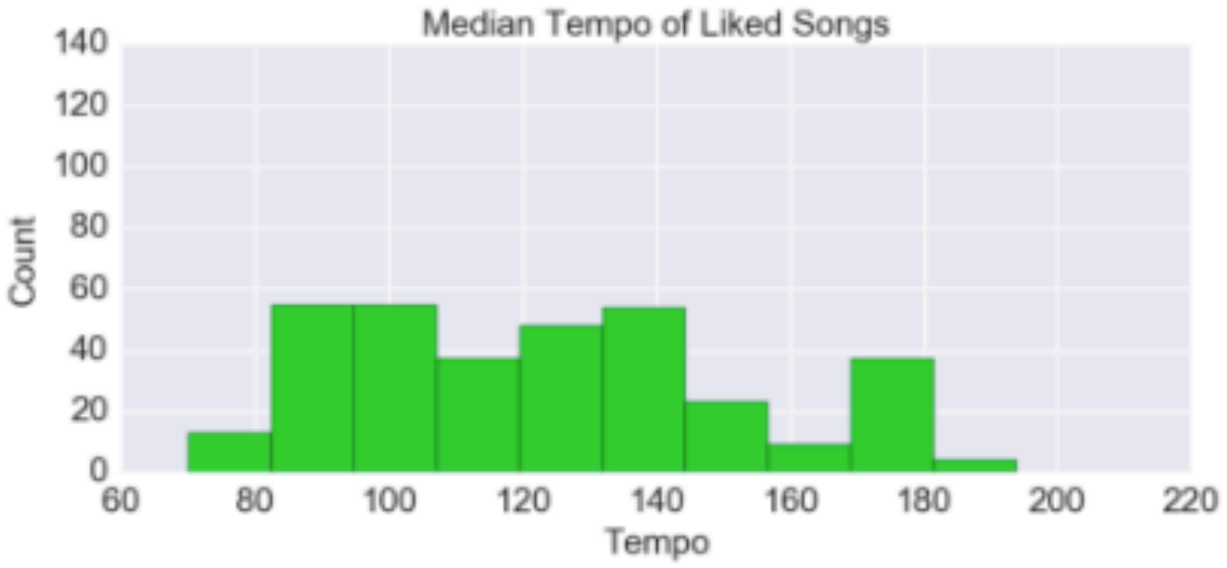
Mean Popularity for Liked Tracks: **32.06**
Mean Popularity for Rejected Tracks: **21.68**
Liked Songs with Popularity equaling zero: **22**
Rejected Songs with Popularity equaling zero: **51**

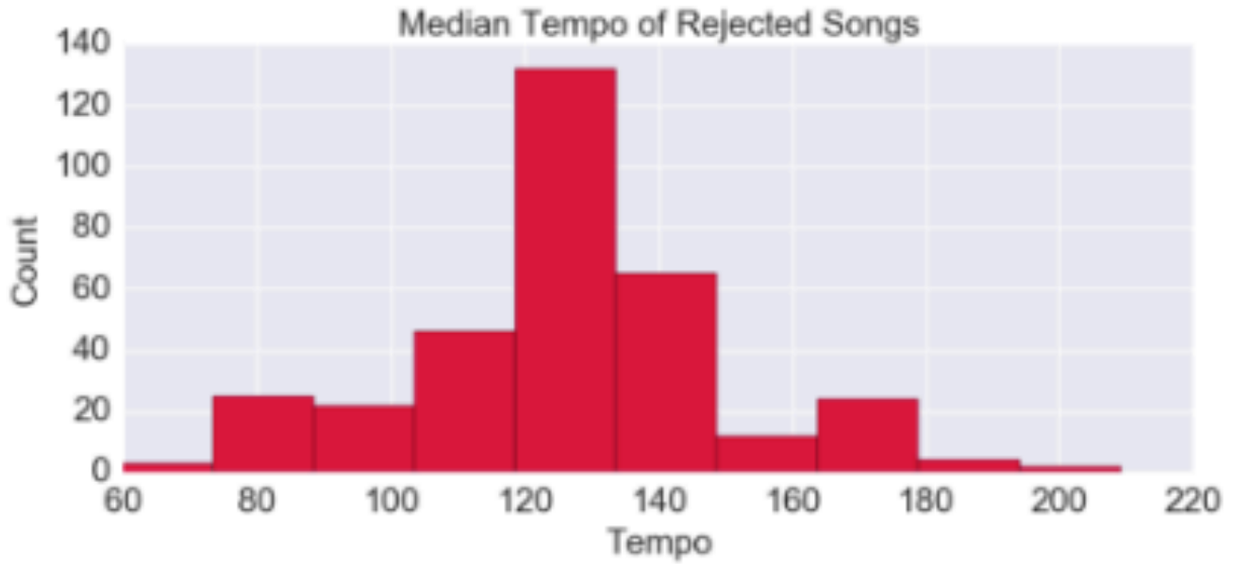

Liked/Rejected Distribution of Popularity
Density
0.025
0.020
0.015
0.010
0.005
0.000
−20
0
20
40
60
80
100
Popularity
Liked
Rejected


Popularity of Liked Songs
Count
0
20
40
60
80
100
0
10
20
30
40
50
60
70
Popularity


Popularity of Rejected Songs
Count
0
20
40
60
80
100
0
10
20
30
40
50
60
70
Popularity


Liked/Rejected Distribution of Median Tempo
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Tempo in BPM
Liked
Rejected

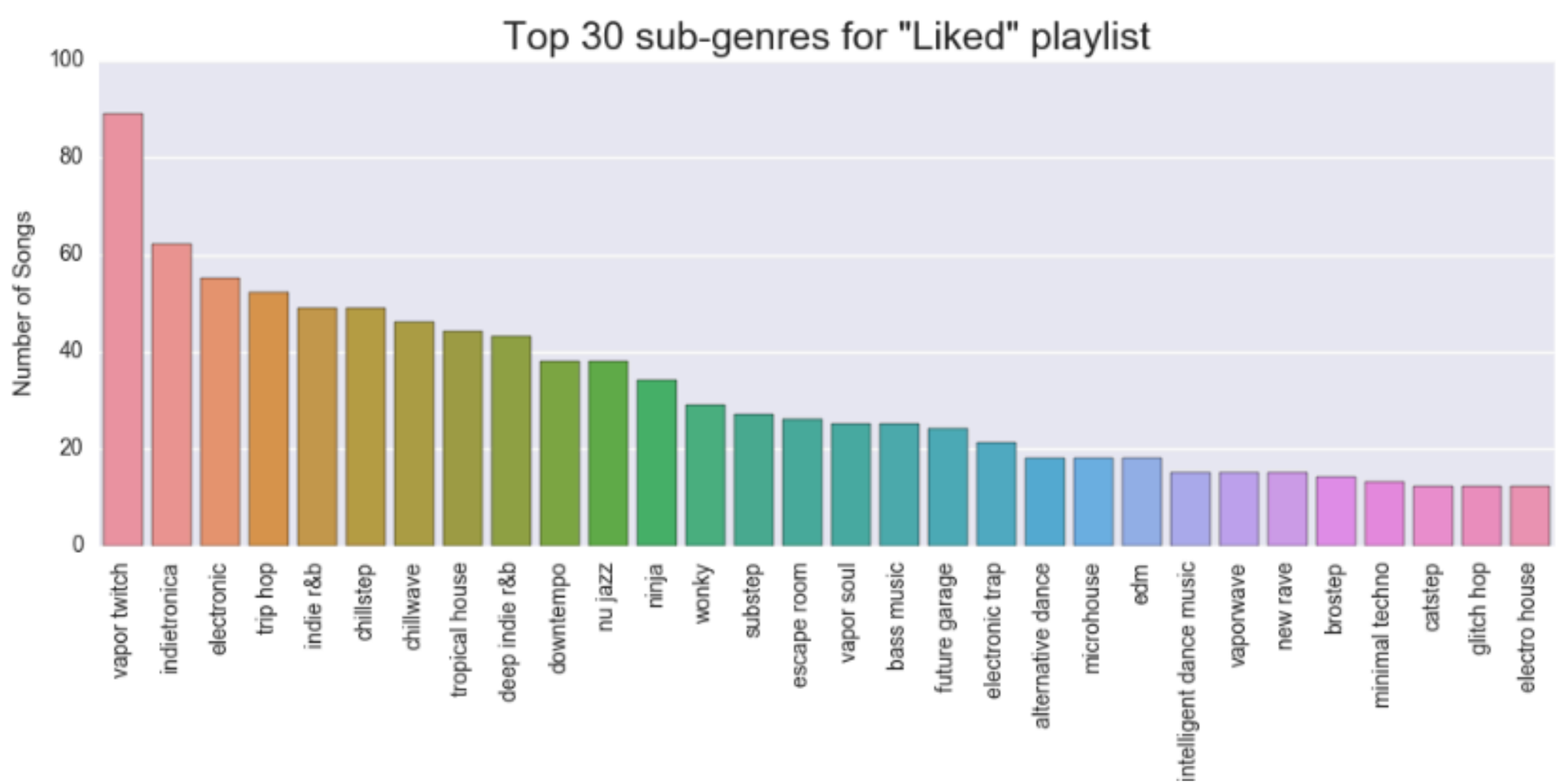

Median Tempo of Liked Songs
Count
140
120
100
80
60
40
20
0
60
80
100
120
140
160
180
200
220
Tempo

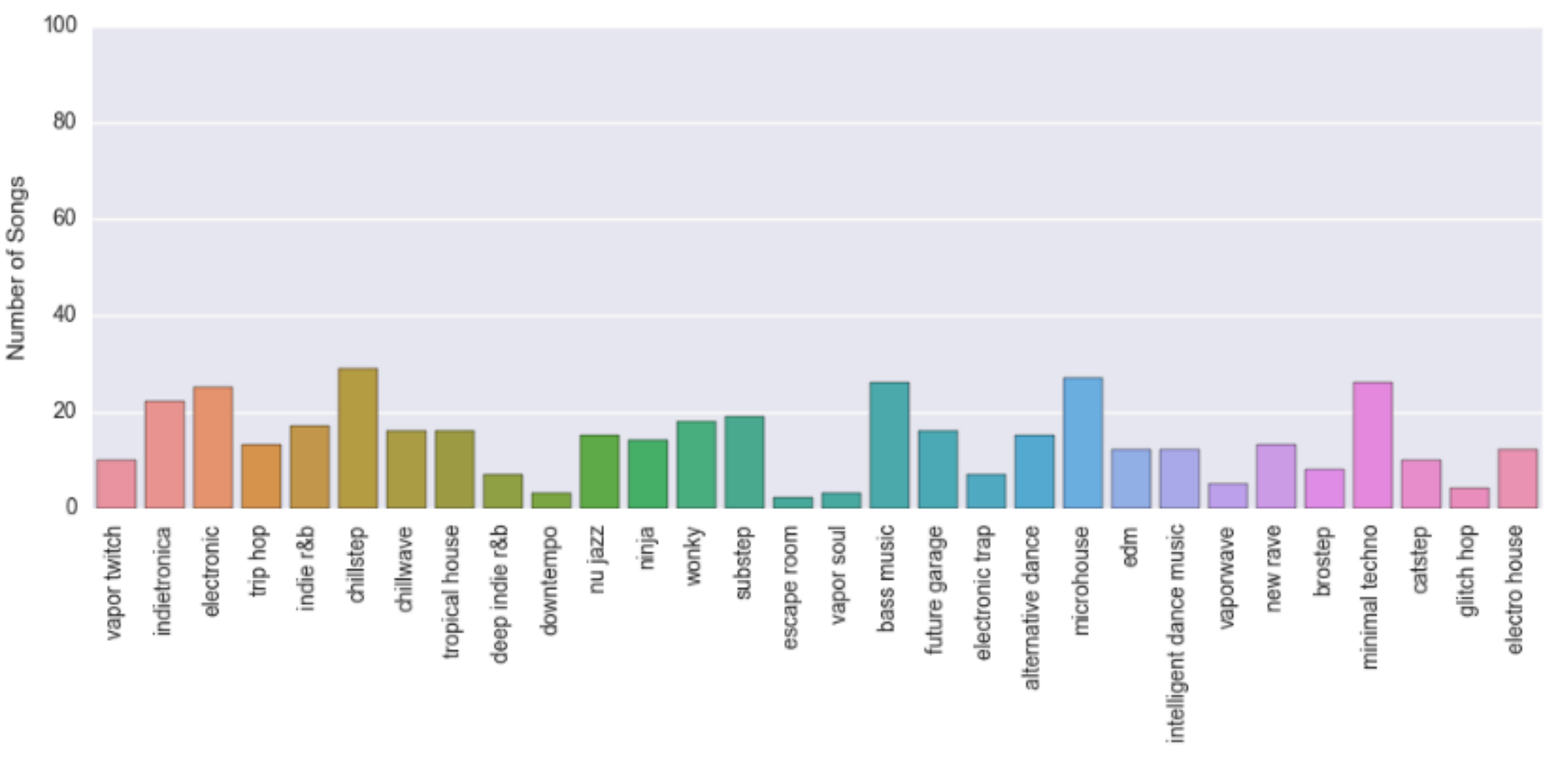

Median Tempo of Rejected Songs
Count
140
120
100
80
60
40
20
0
60
80
100
120
140
160
180
200
220
Tempo

	Song	Artist Genres
10	Takeover	['neurostep', 'vapor twitch']
11	Kissed By A Kisser	['acid jazz', 'ninja', 'nu jazz', 'trip hop']
12	Parks On Fire	['chillstep']
13	Schwindelig - Original	['deep disco house', 'deep euro house', 'deep ...
14	I'll Be Your Reason	['bass trap', 'brostep', 'catstep', 'edm', 'el...
15	Sun Models (feat. Madelyn Grant)	['chillwave', 'edm', 'electronic trap', 'indie...
16	Unfold	['deep tropical house', 'downtempo', 'tropical...
17	Night - Lone Wolf Trait Remix	['bow pop', 'compositional ambient', 'minimal'...
18	Whyarntyou	['bass music', 'chillstep', 'future garage', '...
19	Feeling	['vapor twitch']
20	Ocelot	['chillstep', 'downtempo', 'electronic', 'nu j...

Top 30 Genres in My “Liked” and “Rejected” Playlists

	Liked Sub-Genres	Rejected Sub-Genres
1	(vapor twitch, 89)	(dubstep, 32)
2	(indietronica, 62)	(chillstep, 29)
3	(electronic, 55)	(microhouse, 27)
4	(trip hop, 52)	(bass music, 26)
5	(indie r&b, 49)	(minimal techno, 26)
6	(chillstep, 49)	(electronic, 25)
7	(chillwave, 46)	(house, 24)
8	(tropical house, 44)	(indietronica, 22)
9	(deep indie r&b, 43)	(tech house, 22)
10	(downtempo, 38)	(indie jazz, 21)
11	(nu jazz, 38)	(substep, 19)
12	(ninja, 34)	(fourth world, 18)
13	(wonky, 29)	(wonky, 18)
14	(substep, 27)	(indie r&b, 17)
15	(escape room, 26)	(future garage, 16)
16	(vapor soul, 25)	(tropical house, 16)
17	(bass music, 25)	(chillwave, 16)
18	(future garage, 24)	(alternative dance, 15)
19	(electronic trap, 21)	(nu jazz, 15)
20	(alternative dance, 18)	(ninja, 14)
21	(microhouse, 18)	(techno, 14)
22	(edm, 18)	(trip hop, 13)
23	(intelligent dance music, 15)	(new rave, 13)
24	(vaporwave, 15)	(compositional ambient, 13)
25	(new rave, 15)	(electro house, 12)
26	(brostep, 14)	(intelligent dance music, 12)
27	(minimal techno, 13)	(float house, 12)
28	(catstep, 12)	(minimal tech house, 12)
29	(glitch hop, 12)	(edm, 12)
30	(electro house, 12)	(deep melodic euro house, 11)


Top 30 sub-genres for "Liked" playlist
Number of Songs
0
20
40
60
80
100
vapor twitch
indietronica
electronic
trip hop
indie r&b
chillstep
chillwave
tropical house
deep indie r&b
downtempo
nu jazz
ninja
wonky
substep
escape room
vapor soul
bass music
future garage
electronic trap
alternative dance
microhouse
edm
intelligent dance music
vaporwave
new rave
brostep
minimal techno
catstep
glitch hop
electro house


Number of Songs
0
20
40
60
80
100
vapor twitch
indietronica
electronic
trip hop
indie r&b
chillstep
chillwave
tropical house
deep indie r&b
downtempo
nu jazz
ninja
wonky
substep
escape room
vapor soul
bass music
future garage
electronic trap
alternative dance
microhouse
edm
intelligent dance music
vaporwave
new rave
brostep
minimal techno
catstep
glitch hop
electro house

Model the Data

Common Machine Learning Algorithms

Logistic Regression

Naive Bayes

Support Vector Machine

Decision Tree

K-Nearest Neighbor

K-Means

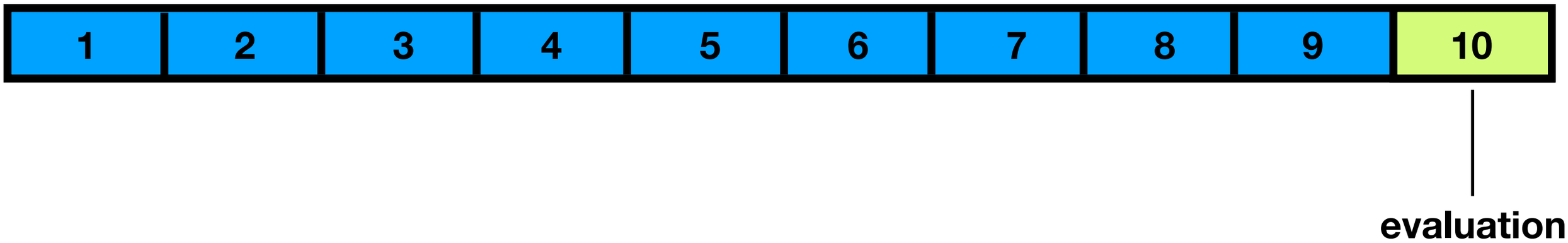
Neural Network

DATA SET

DATA SET

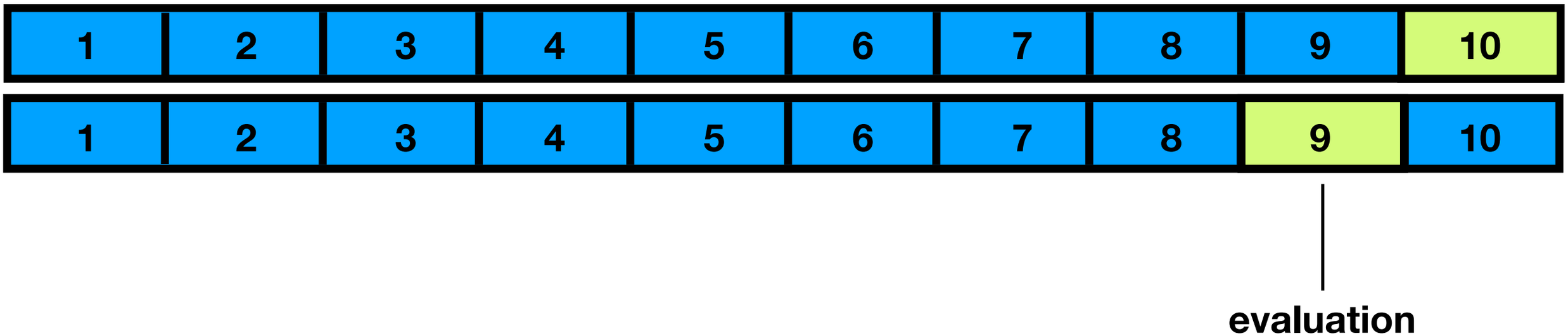
10-Fold Cross Validation

Training Set



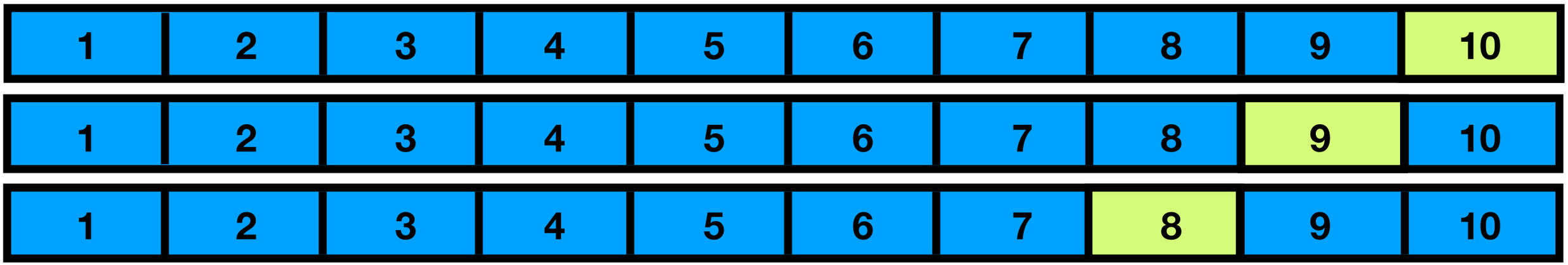
10-Fold Cross Validation

Training Set



10-Fold Cross Validation

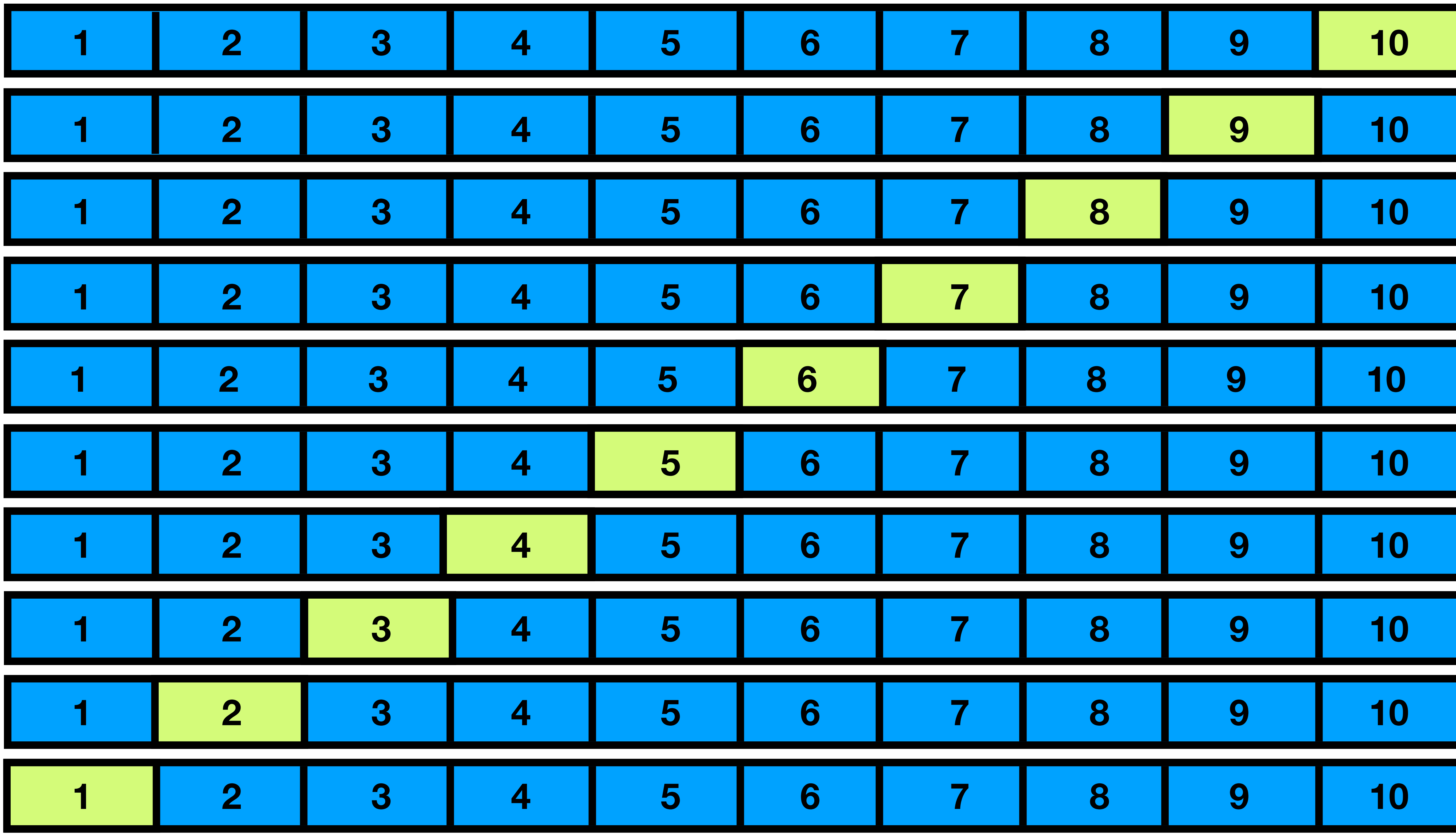
Training Set



evaluation

10-Fold Cross Validation

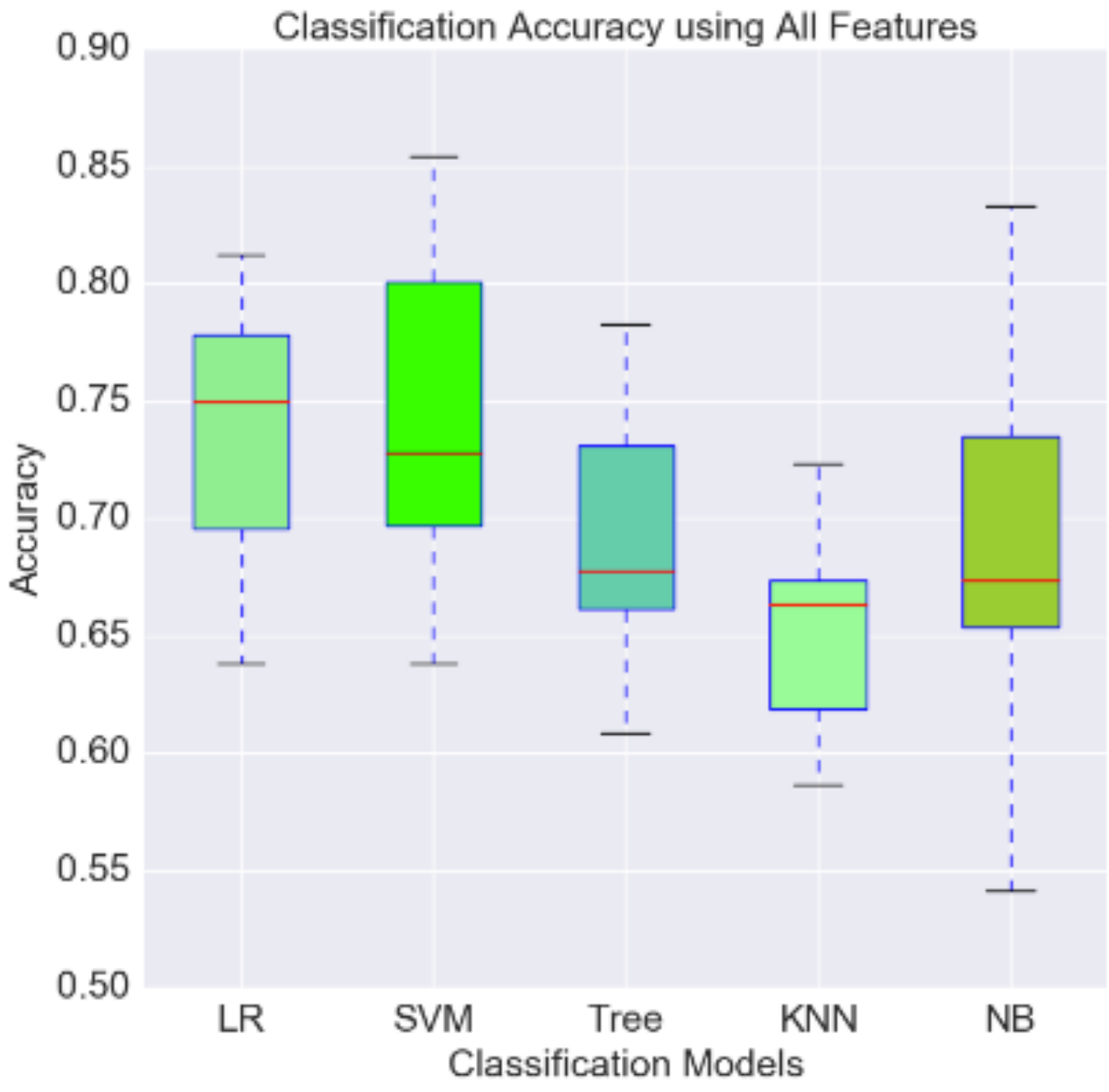
Training Set



Logistic Regression

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	mean
0.69	0.64	0.73	0.82	0.64	0.70	0.68	0.71	0.70	0.69	**0.70**

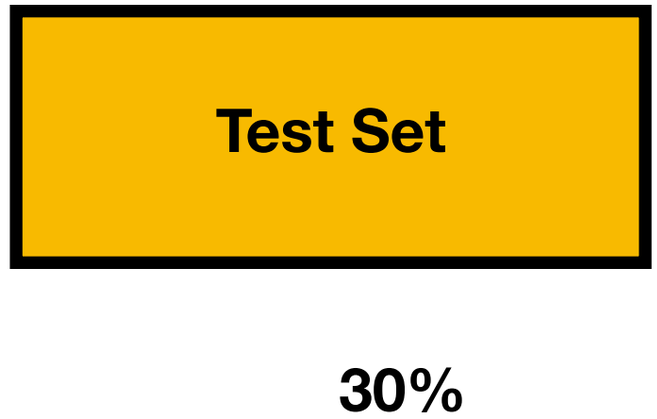
Logistic Regression	Support Vector Machine	Decision Tree	K-Nearest Neighbor	Naive Bayes
0.705	0.722	0.635	0.675	0.607

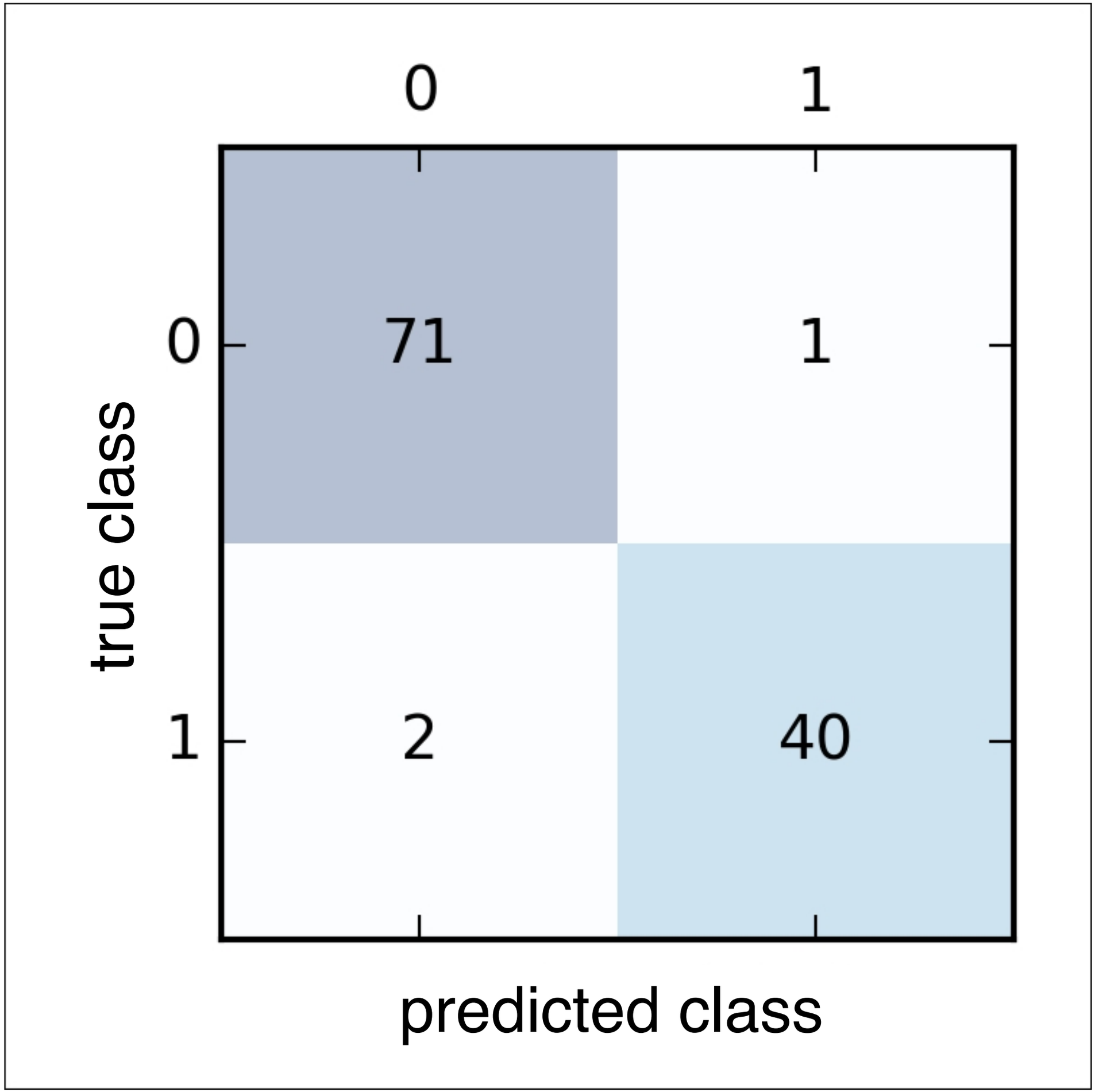


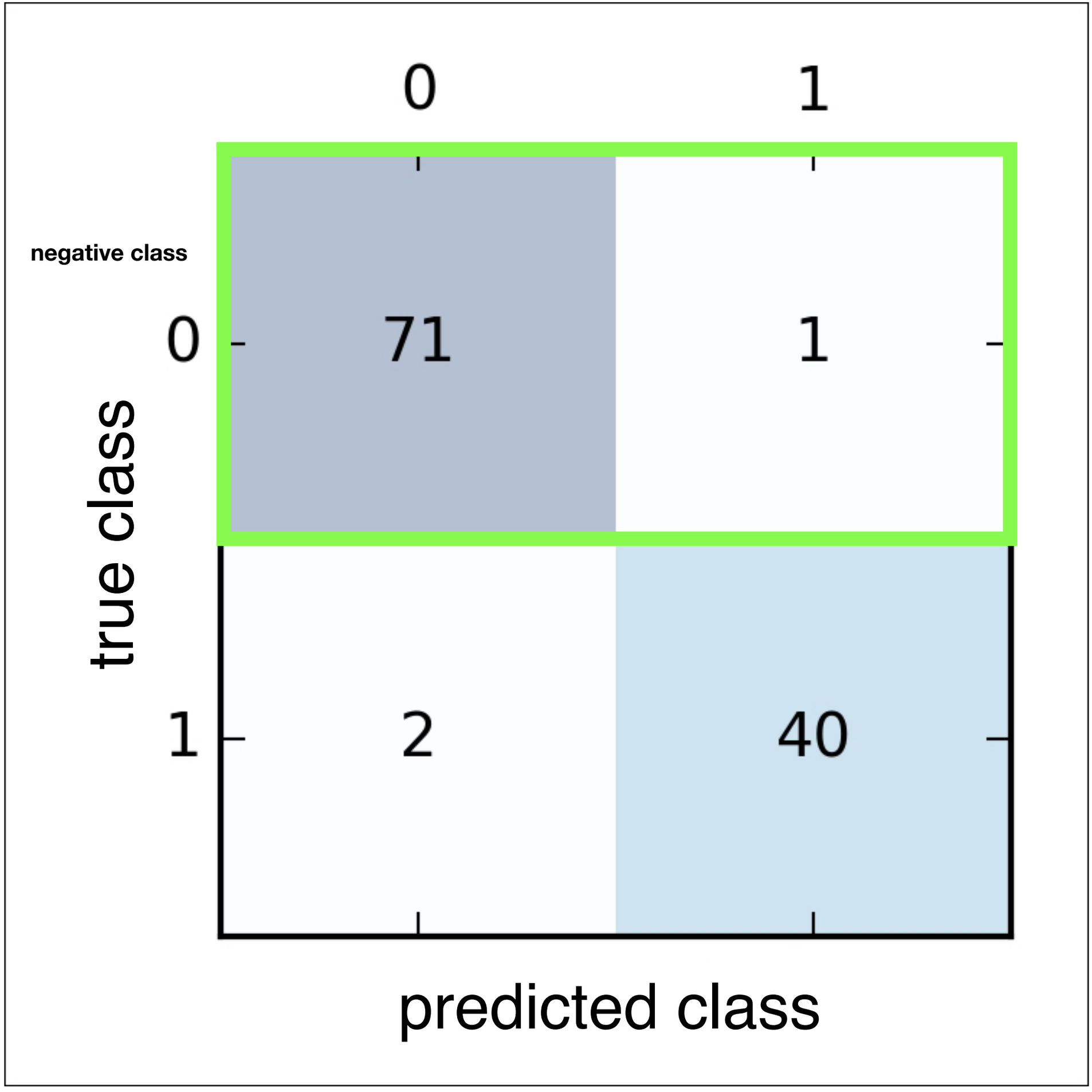
Training Set

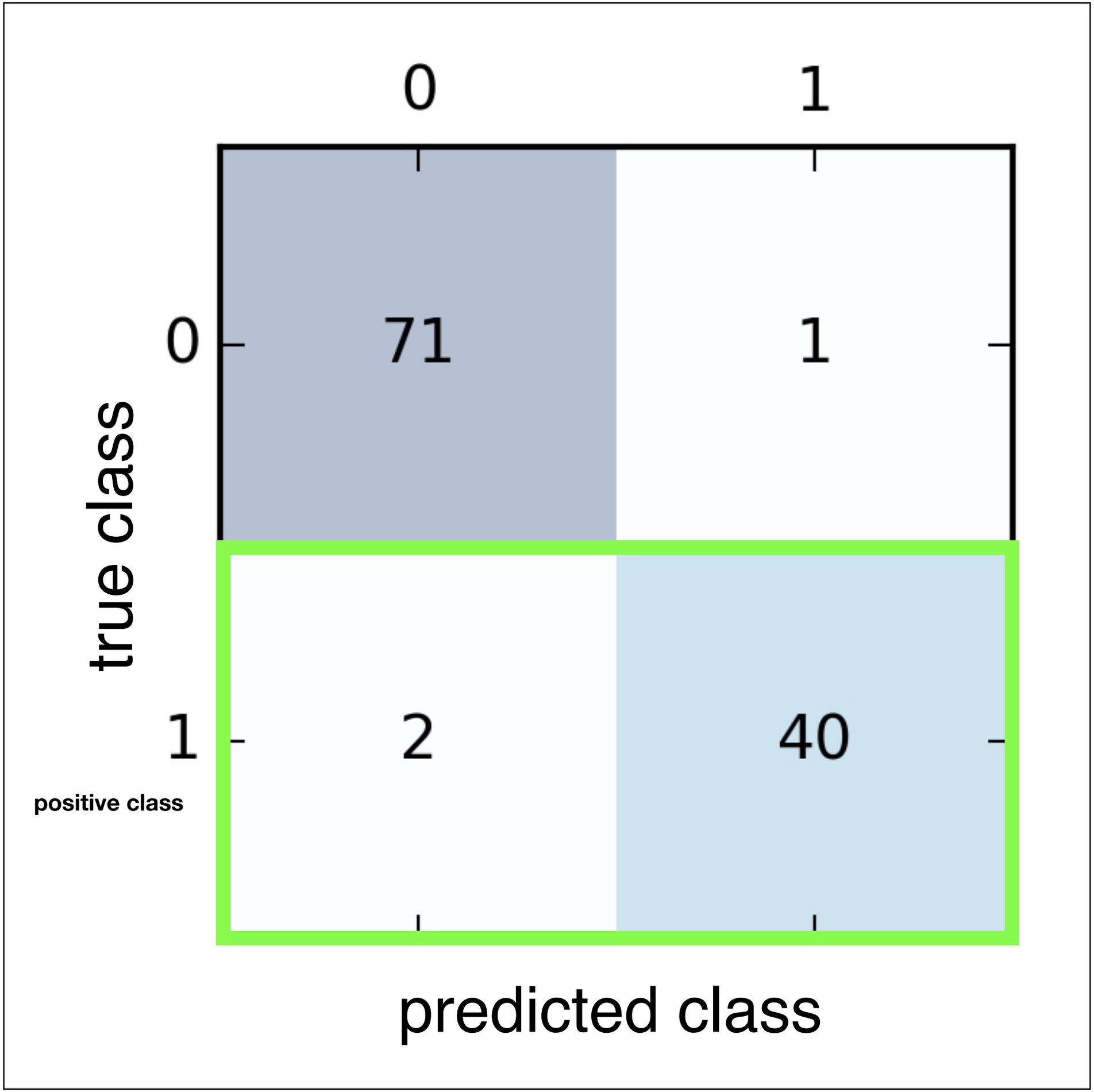
DATA SET

DATA SET




0
1
0
71
1
true class
1
2
40
predicted class


0
1
negative class
0
71
1
true class
1
2
40
predicted class


0
1
0
71
1
true class
1
2
40
positive class
predicted class

Confusion Matrix

Negative Class	**TN** True Negative	**FP** False Positive
Positive Class	**FN** False Negative	**TP** True Positive
	Predicted Negative	Predicted Positive

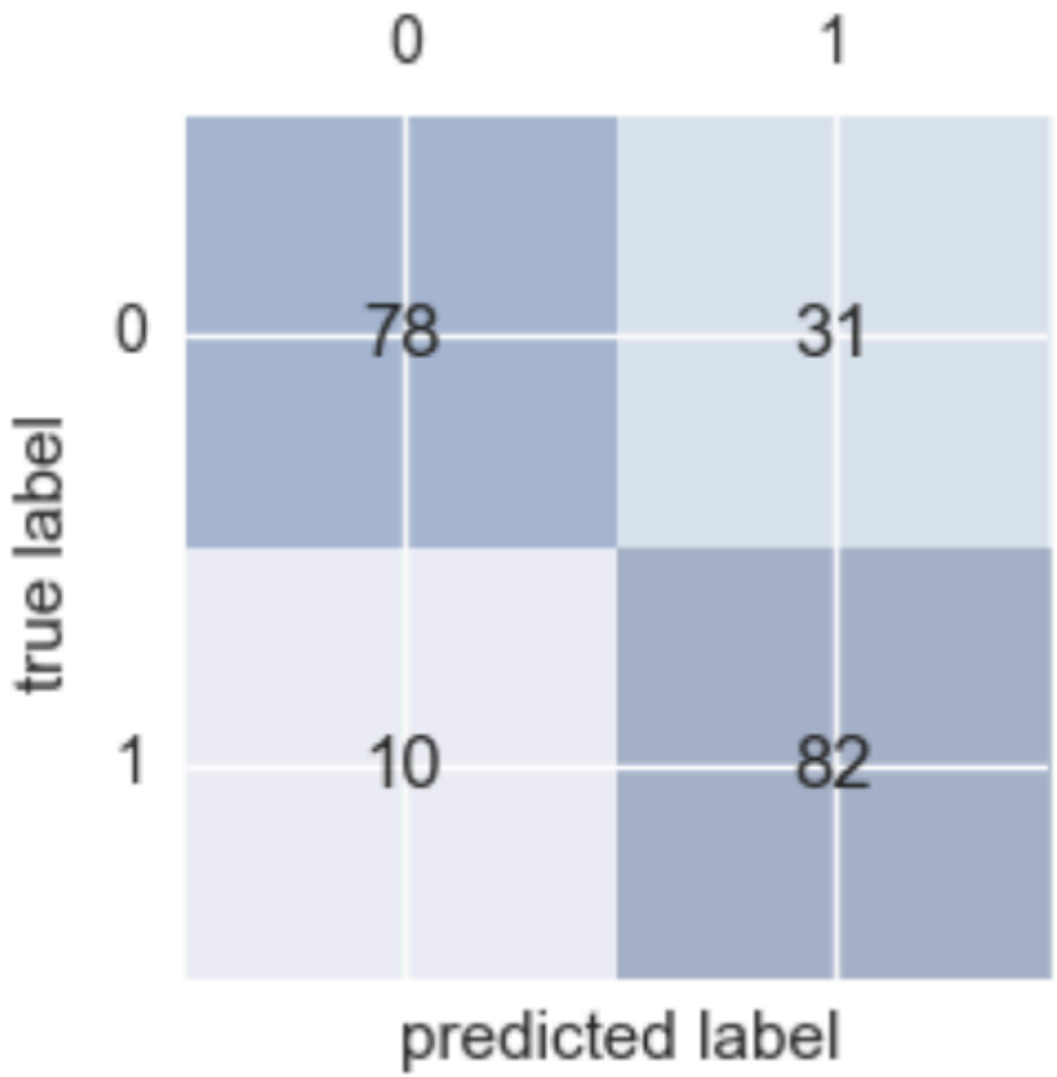
Model Evaluation Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

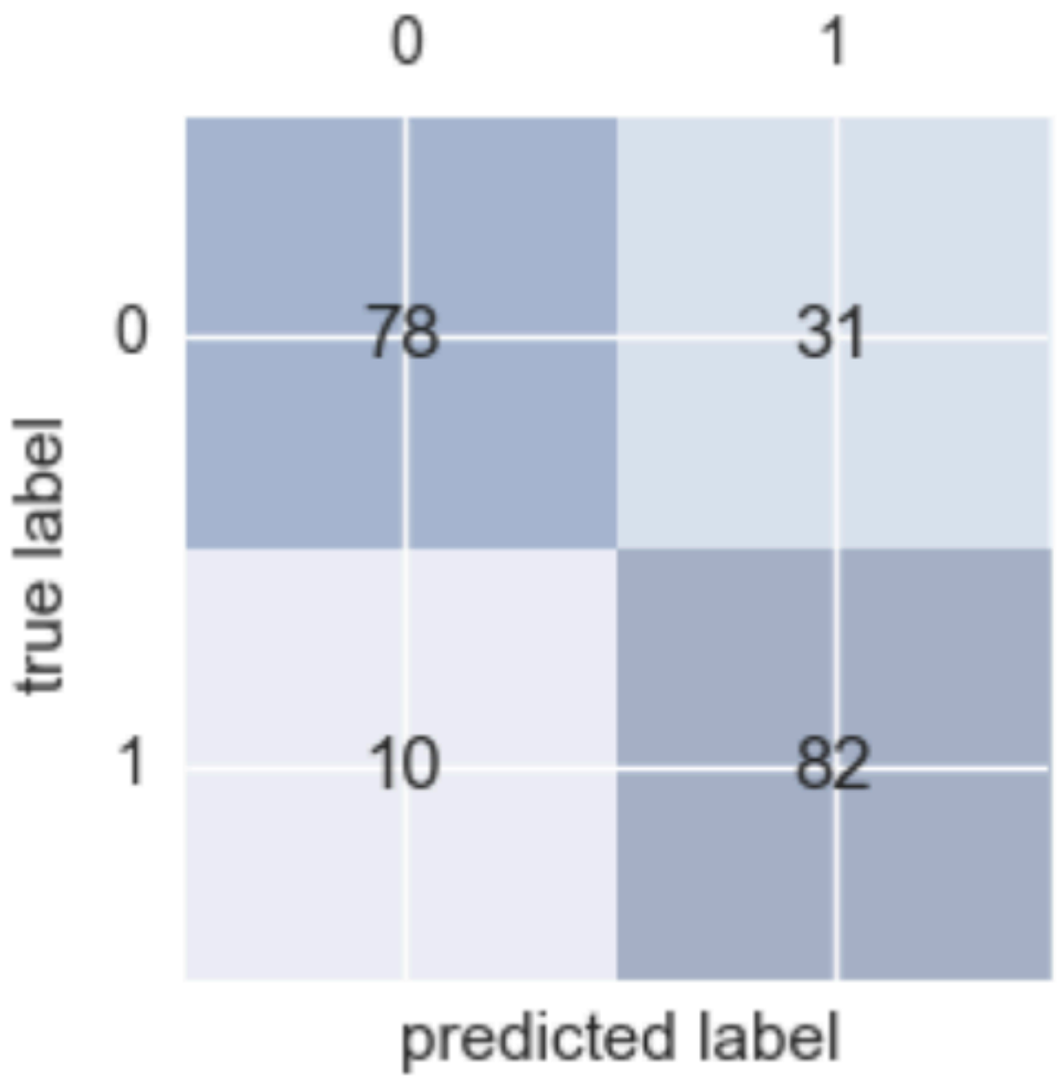
Spotify “Liked” Classification



Accuracy
0.796

Spotify “Liked” Classification

	Precision	Recall	F1-score
Rejected	0.89	0.72	0.79
Liked	0.73	0.80	0.80



DATA SET

100%

Useful Machine Learning and Data Science books for beginners to intermediate:

Book	Author
An Introduction to Statistical Learning	Robert Tibshirani and Trevor Hastie
Python Machine Learning	Sebastian Raschka
Introduction to Machine Learning with Python	Andreas C. Müller; Sarah Guido
Data Smart: Using Data Science to Transform Information Into Insight	John W. Foreman
Naked Statistics: Stripping the Dread from the Data	Charles Wheelan