Udacity Machine Learning Capstone Project: August 2020

Greyhound Race Prediction

Tony Ward

Definition

Project overview

Greyhound racing is an organized, competitive sport in which greyhounds are raced around a track. As with horse racing, greyhound races often allow the public to bet on the outcome.[1]

I was introduced to betting on greyhounds whilst at University. About once a week a group of us would head down to the local book maker and bet a few pounds, which provided entertainment and way of winding down between lecturers. I remember checking the form guide and comparing each's dogs previous winning times which soon became overwhelming. I realised at that time there was an opportunity to use the past racing results to guide the betting decision in an intelligent way.

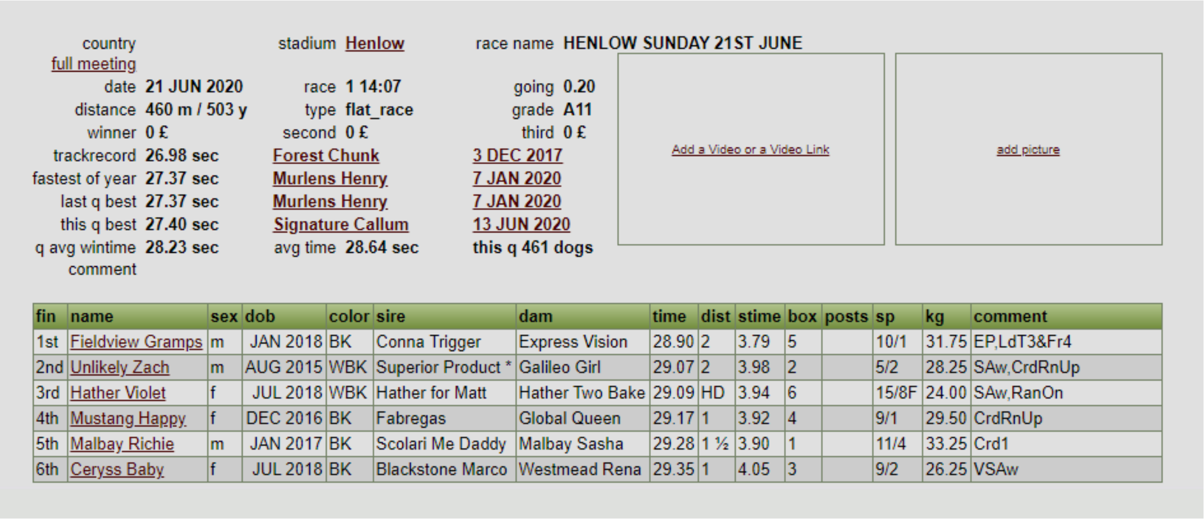
Problem Statement

This project is focused on using historic data on greyhound races to investigate whether a machine learning system can accurately predict the results of the UK greyhound racing scene.

Dataset and Inputs

http://www.greyhound-data.com/ is website which provides information about greyhounds from all over the world with pedigree information drawn from the last four centuries. Online are 4,549,034 race results and 2,333,577 greyhound pedigrees.

Here is an example of the data available for a given race



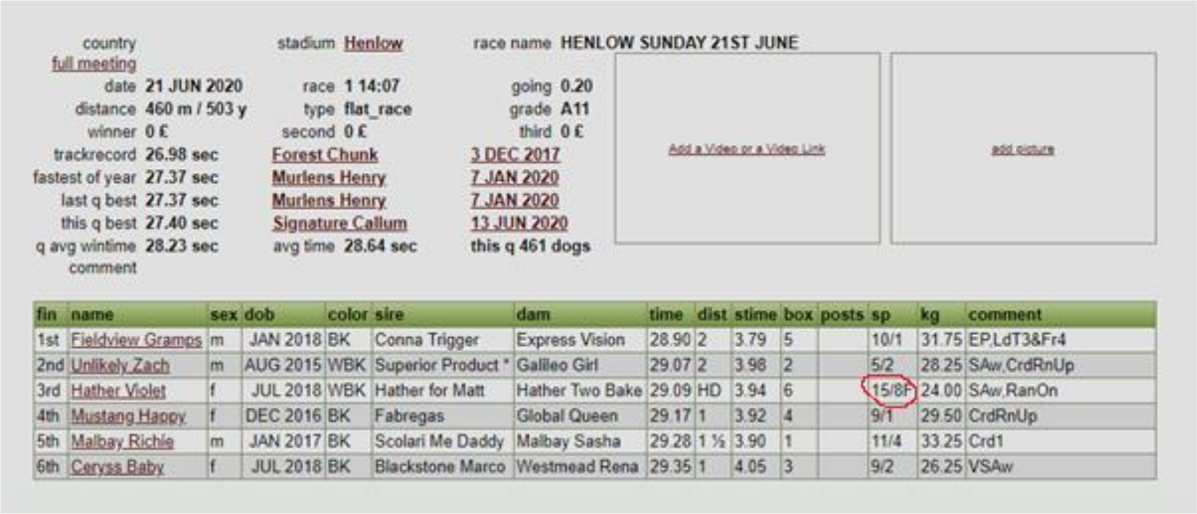
country		stadium	Henlow	race name	HENLOW SUNDAY 21ST JUNE
full meeting					
date	21 JUN 2020	race	1 14:07	going	0.20
distance	460 m / 503 y	type	flat_race	grade	A11
winner	0 £	second	0 £	third	0 £
trackrecord	26.98 sec	Forest Chunk		3 DEC 2017	
fastest of year	27.37 sec	Murlens Henry		7 JAN 2020	
last q best	27.37 sec	Murlens Henry		7 JAN 2020	
this q best	27.40 sec	Signature Callum		13 JUN 2020	
q avg wintime	28.23 sec	avg time	28.64 sec	this q 461 dogs	
comment					

fin	name	sex	dob	color	sire	dam	time	dist	stime	box	posts	sp	kg	comment
1st	Fieldview Gramps	m	JAN 2018	BK	Conna Trigger	Express Vision	28.90	2	3.79	5		10/1	31.75	EP,LdT3&Fr4
2nd	Unlikely Zach	m	AUG 2015	WBK	Superior Product *	Galileo Girl	29.07	2	3.98	2		5/2	28.25	SAw,CrdRnUp
3rd	Hather Violet	f	JUL 2018	WBK	Hather for Matt	Hather Two Bake	29.09	HD	3.94	6		15/8F	24.00	SAw,RanOn
4th	Mustang Happy	f	DEC 2016	BK	Fabregas	Global Queen	29.17	1	3.92	4		9/1	29.50	CrdRnUp
5th	Malbay Richie	m	JAN 2017	BK	Scolari Me Daddy	Malbay Sasha	29.28	1 ½	3.90	1		11/4	33.25	Crd1
6th	Ceryss Baby	f	JUL 2018	BK	Blackstone Marco	Westmead Rena	29.35	1	4.05	3		9/2	26.25	VSAw

[1] https://en.wikipedia.org/wiki/Greyhound_racing

Solution

As part of this project I scraped 18 years' worth of race data for one stadium (Monmore), which constitutes approximately 55k races. Next, I designed a Postgres SQL database and inserted the data for later analysis. Using SQL queries I was able to construct several features which were later used to build and validate a machine learning model whose goal it was to predict the eventual race winner. This model has been evaluated against the benchmark given by the bookmakers favourite, which is indicated as either a F (favourite) or JF (joint favourite) in the sp (starting price) column



country
full meeting
date 21 JUN 2020
distance 460 m / 503 y
winner 0 £
trackrecord 26.98 sec
fastest of year 27.37 sec
last q best 27.37 sec
this q best 27.40 sec
q avg wintime 28.23 sec
comment

stadium Henlow
race 1 14:07
type flat_race
second 0 £
Forest Chunk
Murlens Henry
Murlens Henry
Signature Callum
avg time 28.64 sec

race name HENLOW SUNDAY 21ST JUNE
going 0.20
grade A11
third 0 £
3 DEC 2017
7 JAN 2020
7 JAN 2020
13 JUN 2020
this q 461 dogs

fin	name	sex	dob	color	sire	dam	time	dist	stime	box	posts	sp	kg	comment
1st	Fieldview Gramps	m	JAN 2018	BK	Conna Trigger	Express Vision	28.90	2	3.79	5		10/1	31.75	EP,LdT3&Fr4
2nd	Unlikely Zach	m	AUG 2015	WBK	Superior Product *	Galileo Girl	29.07	2	3.98	2		5/2	28.25	SAw,CrdRnUp
3rd	Hather Violet	f	JUL 2018	WBK	Hather for Matt	Hather Two Bake	29.09	HD	3.94	6		15/8F	24.00	SAw,RanOn
4th	Mustang Happy	f	DEC 2016	BK	Fabregas	Global Queen	29.17	1	3.92	4		9/1	29.50	CrdRnUp
5th	Malbay Richie	m	JAN 2017	BK	Scolari Me Daddy	Malbay Sasha	29.28	1 ½	3.90	1		11/4	33.25	Crd1
6th	Ceryss Baby	f	JUL 2018	BK	Blackstone Marco	Westmead Rena	29.35	1	4.05	3		9/2	26.25	VSAw

Metrics

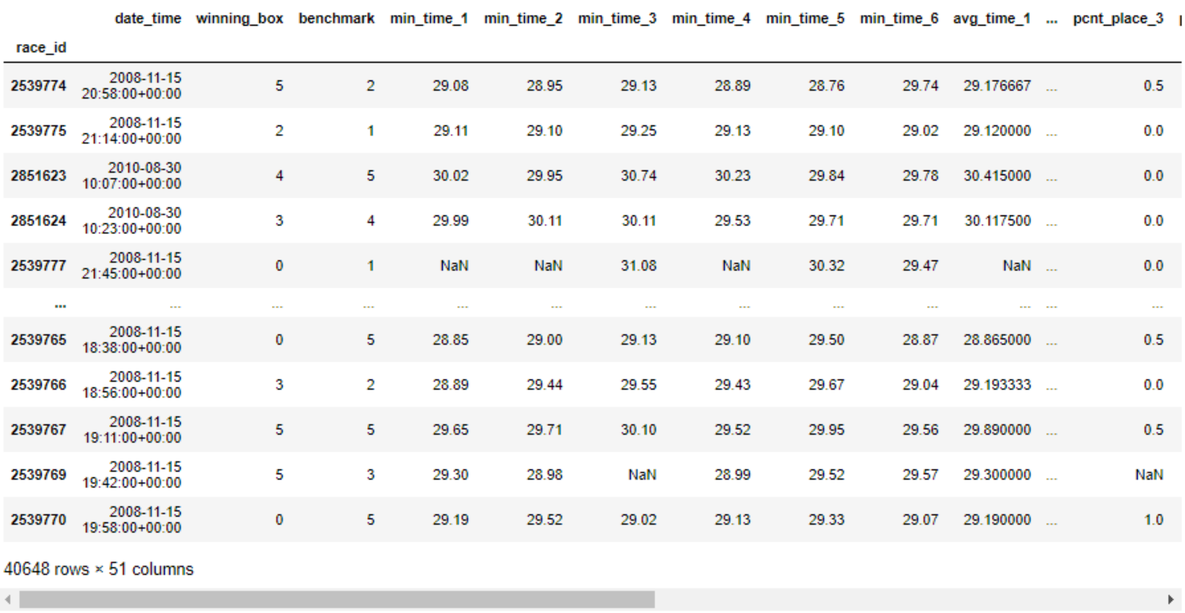
For each race we must predict a race winner. This prediction will be compared to the eventual race winner.

Greyhound	Predicted Winner	Actual Winner
1	0	0
2	0	0
3	0	1
4	1	0
5	0	0
6	0	0

To evaluate the model against the benchmark we will use the accuracy metric. This will tell us the proportion of races where we correctly forecasted the race winner.

Framing the Problem

We treat this as a multi class classification problem where we must predict the box (1 to 6) that the winning dog started the race from. We will organise our data such that each row represents an individual race, with features relating to each of the 6 dogs running in the race. For example



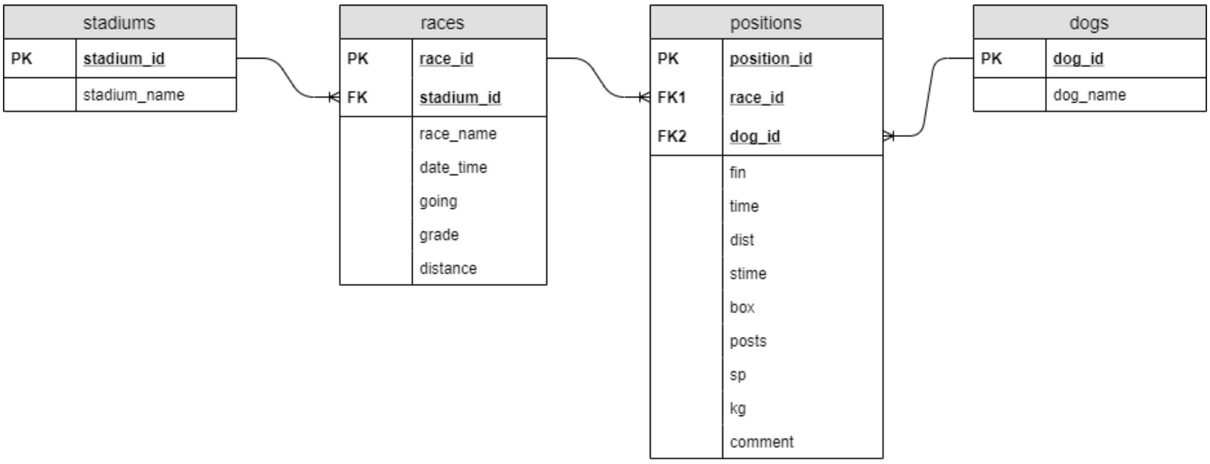
race_id	date_time	winning_box	benchmark	min_time_1	min_time_2	min_time_3	min_time_4	min_time_5	min_time_6	avg_time_1	...	pcnt_place_3
2539774	2008-11-15 20:58:00+00:00	5	2	29.08	28.95	29.13	28.89	28.76	29.74	29.176667	...	0.5
2539775	2008-11-15 21:14:00+00:00	2	1	29.11	29.10	29.25	29.13	29.10	29.02	29.120000	...	0.0
2851623	2010-08-30 10:07:00+00:00	4	5	30.02	29.95	30.74	30.23	29.84	29.78	30.415000	...	0.0
2851624	2010-08-30 10:23:00+00:00	3	4	29.99	30.11	30.11	29.53	29.71	29.71	30.117500	...	0.0
2539777	2008-11-15 21:45:00+00:00	0	1	NaN	NaN	31.08	NaN	30.32	29.47	NaN	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
2539765	2008-11-15 18:38:00+00:00	0	5	28.85	29.00	29.13	29.10	29.50	28.87	28.865000	...	0.5
2539766	2008-11-15 18:56:00+00:00	3	2	28.89	29.44	29.55	29.43	29.67	29.04	29.193333	...	0.0
2539767	2008-11-15 19:11:00+00:00	5	5	29.65	29.71	30.10	29.52	29.95	29.56	29.890000	...	0.5
2539769	2008-11-15 19:42:00+00:00	5	3	29.30	28.98	NaN	28.99	29.52	29.57	29.300000	...	NaN
2539770	2008-11-15 19:58:00+00:00	0	5	29.19	29.52	29.02	29.13	29.33	29.07	29.190000	...	1.0

40648 rows × 51 columns

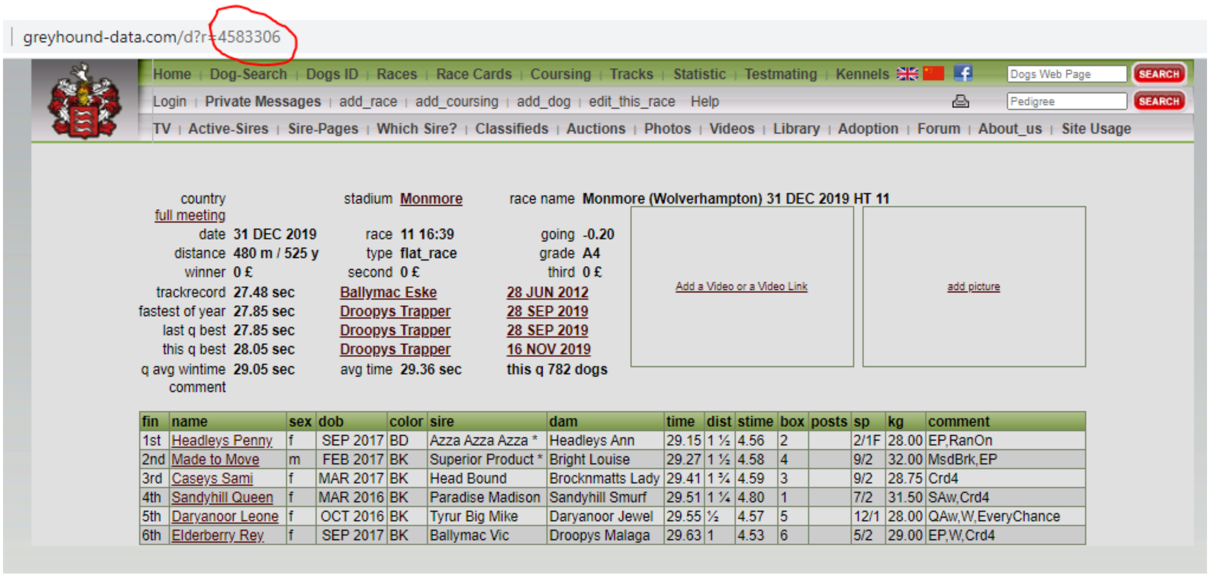
Rows are indexed by race_id which is a unique identifier for each race. The winning_box is the target variable and the benchmark column relates to the bookmakers favourite. Note how that both these columns are indexed to start at zero rather than one – this is purely because lightgbm requires the data this way. Min_time_1 is the fastest time run by the dog starting from box 1 in the last 25 days.

Data Exploration and Visualisation

Data is stored in a Postgres database under the following schema



A typical race contains several pieces of information. An example race is provided below

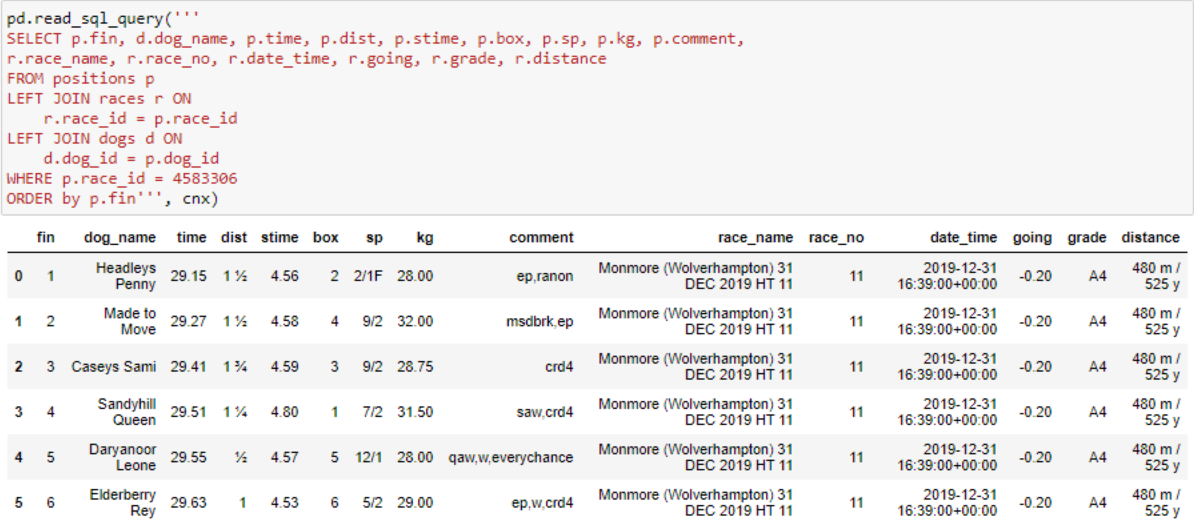

greyhound-data.com/d?r=4583306
Home | Dog-Search | Dogs ID | Races | Race Cards | Coursing | Tracks | Statistic | Testmating | Kennels
Dogs Web Page
SEARCH
Login | Private Messages | add_race | add_coursing | add_dog | edit_this_race Help
Pedigree
SEARCH
TV | Active-Sires | Sire-Pages | Which Sire? | Classifieds | Auctions | Photos | Videos | Library | Adoption | Forum | About_us | Site Usage
Add a Video or a Video Link
add picture

country		stadium	Monmore	race name	Monmore (Wolverhampton) 31 DEC 2019 HT 11
full meeting					
date	31 DEC 2019	race	11 16:39	going	-0.20
distance	480 m / 525 y	type	flat_race	grade	A4
winner	0 £	second	0 £	third	0 £
trackrecord	27.48 sec	Ballymac Eske		28 JUN 2012	
fastest of year	27.85 sec	Droopys Trapper		28 SEP 2019	
last q best	27.85 sec	Droopys Trapper		28 SEP 2019	
this q best	28.05 sec	Droopys Trapper		16 NOV 2019	
q avg wintime	29.05 sec	avg time	29.36 sec	this q 782 dogs	
comment					

fin	name	sex	dob	color	sire	dam	time	dist	stime	box	posts	sp	kg	comment
1st	Headleys Penny	f	SEP 2017	BD	Azza Azza Azza *	Headleys Ann	29.15	1 ½	4.56	2		2/1F	28.00	EP,RanOn
2nd	Made to Move	m	FEB 2017	BK	Superior Product *	Bright Louise	29.27	1 ½	4.58	4		9/2	32.00	MsdBrk,EP
3rd	Caseys Sami	f	MAR 2017	BK	Head Bound	Brocknmatts Lady	29.41	1 ¾	4.59	3		9/2	28.75	Crd4
4th	Sandyhill Queen	f	MAR 2016	BK	Paradise Madison	Sandyhill Smurf	29.51	1 ¼	4.80	1		7/2	31.50	SAw,Crd4
5th	Daryanoor Leone	f	OCT 2016	BK	Tyrur Big Mike	Daryanoor Jewel	29.55	½	4.57	5		12/1	28.00	QAw,W,EveryChance
6th	Elderberry Rey	f	SEP 2017	BK	Ballymac Vic	Droopys Malaga	29.63	1	4.53	6		5/2	29.00	EP,W,Crd4

The URL contains a unique identifier which we call race_id. The race_id is the primary key in both the races and positions tables.

To avoid duplication and protect the integrity of the relationships we have chosen to store the data in third normal form. We can reconstruct the above table using the following query (note that due to time constraints we did not capture the sex, dob, colour, sire, dam information of the dog)

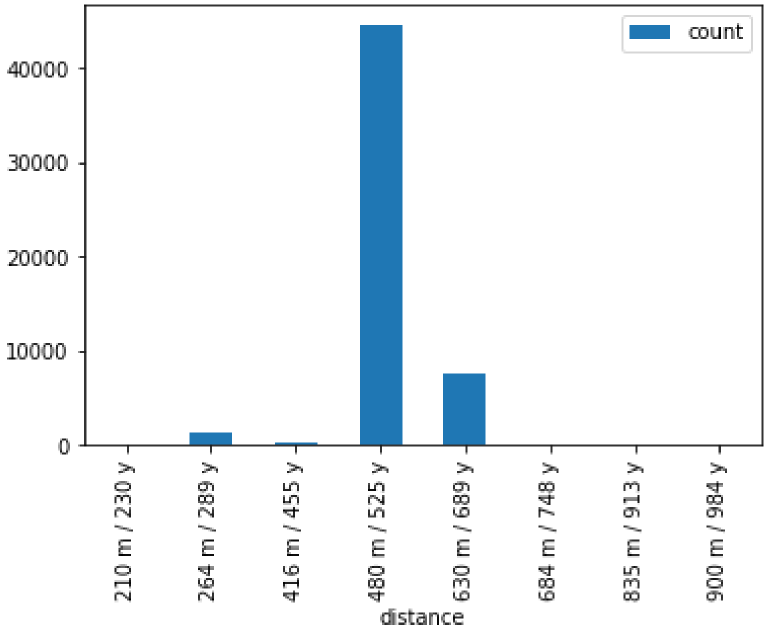


```
pd.read_sql_query('''
SELECT p.fin, d.dog_name, p.time, p.dist, p.stime, p.box, p.sp, p.kg, p.comment,
r.race_name, r.race_no, r.date_time, r.going, r.grade, r.distance
FROM positions p
LEFT JOIN races r ON
    r.race_id = p.race_id
LEFT JOIN dogs d ON
    d.dog_id = p.dog_id
WHERE p.race_id = 4583306
ORDER by p.fin''', cnx)
```

	fin	dog_name	time	dist	stime	box	sp	kg	comment	race_name	race_no	date_time	going	grade	distance
0	1	Headleys Penny	29.15	1 ½	4.56	2	2/1F	28.00	ep,ranon	Monmore (Wolverhampton) 31 DEC 2019 HT 11	11	2019-12-31 16:39:00+00:00	-0.20	A4	480 m / 525 y
1	2	Made to Move	29.27	1 ½	4.58	4	9/2	32.00	msdbrk,ep	Monmore (Wolverhampton) 31 DEC 2019 HT 11	11	2019-12-31 16:39:00+00:00	-0.20	A4	480 m / 525 y
2	3	Caseys Sami	29.41	1 ¾	4.59	3	9/2	28.75	crd4	Monmore (Wolverhampton) 31 DEC 2019 HT 11	11	2019-12-31 16:39:00+00:00	-0.20	A4	480 m / 525 y
3	4	Sandyhill Queen	29.51	1 ¼	4.80	1	7/2	31.50	saw,crd4	Monmore (Wolverhampton) 31 DEC 2019 HT 11	11	2019-12-31 16:39:00+00:00	-0.20	A4	480 m / 525 y
4	5	Daryanoor Leone	29.55	½	4.57	5	12/1	28.00	qaw,w,everychance	Monmore (Wolverhampton) 31 DEC 2019 HT 11	11	2019-12-31 16:39:00+00:00	-0.20	A4	480 m / 525 y
5	6	Elderberry Rey	29.63	1	4.53	6	5/2	29.00	ep,w,crd4	Monmore (Wolverhampton) 31 DEC 2019 HT 11	11	2019-12-31 16:39:00+00:00	-0.20	A4	480 m / 525 y

Distance

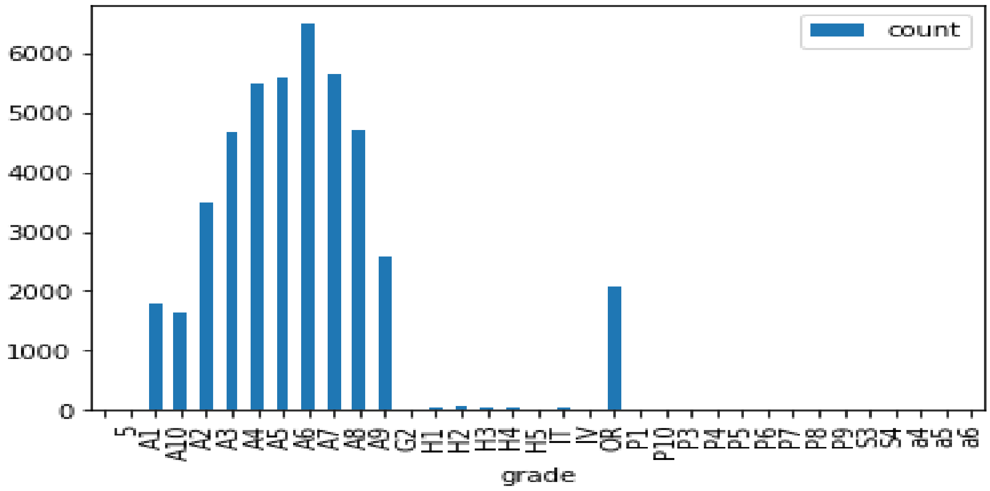
Dogs race over a range of distances



For simplicity we will restrict ourselves to the most frequently occurring race distance of 480m.

Grade

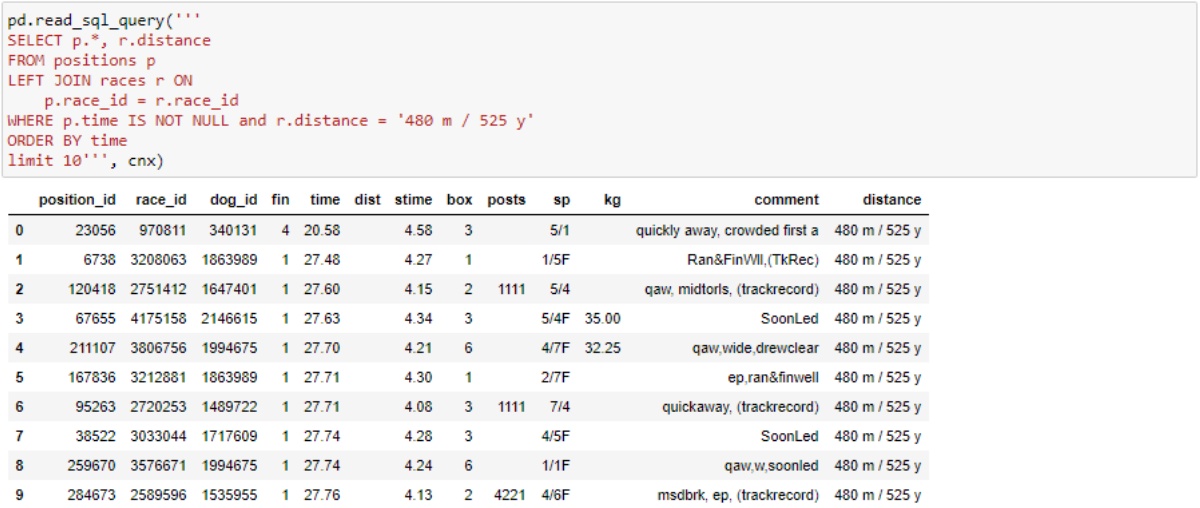
Races are organised in such a way so that dogs of similar ability race each other. The most competitive grade that has the fastest dogs are Open Races (OR). In these races' dogs travel across the country and race against dogs from different stadiums. Grades A1 to A10 are local races where the dogs will race at the same stadium each week. A1 has the fastest dogs and A10 the slowest. Dogs are moved up and down a grade at the discretion of the race director based on recent form. A dog is automatically moved up a grade if it wins a race, and moves down a grade if it fails to place (finish in the top 3) for three races in a row



We will only include races of grade A1 – A10 in our modelling, and exclude Open Races and everything else. It is hoped that there is more chance of successfully predicting races where the dogs run on the same track.

Time

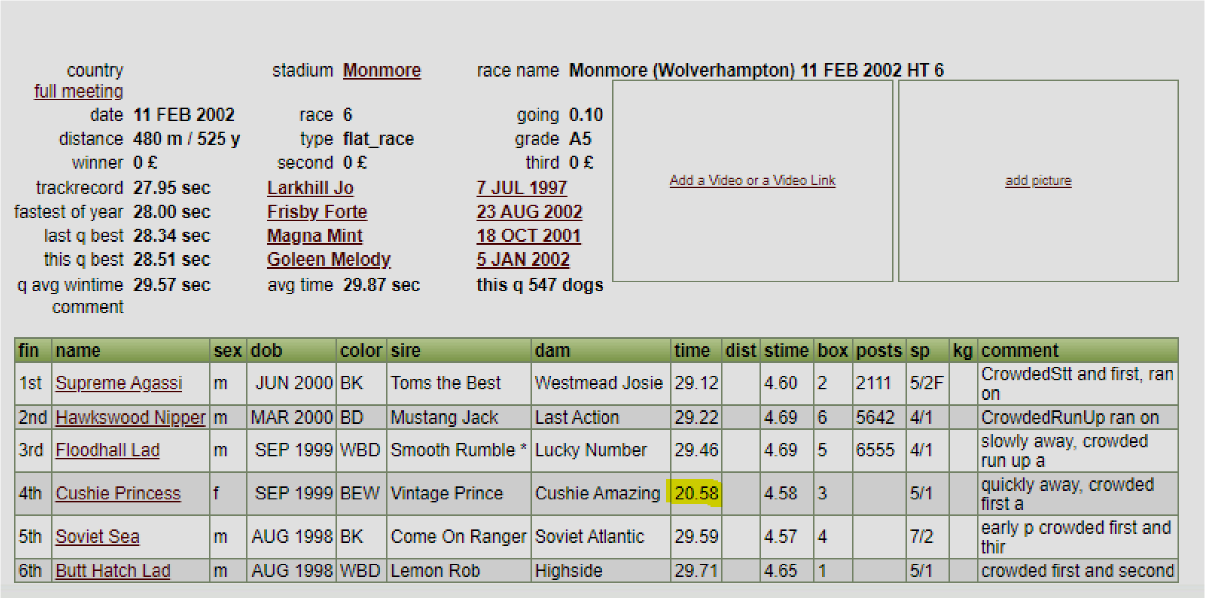
Lets have a look at the fastest race times for the chosen race distance of 480m



```
pd.read_sql_query('''
SELECT p.*, r.distance
FROM positions p
LEFT JOIN races r ON
    p.race_id = r.race_id
WHERE p.time IS NOT NULL and r.distance = '480 m / 525 y'
ORDER BY time
limit 10''', cnx)
```

	position_id	race_id	dog_id	fin	time	dist	stime	box	posts	sp	kg	comment	distance
0	23056	970811	340131	4	20.58		4.58	3		5/1		quickly away, crowded first a	480 m / 525 y
1	6738	3208063	1863989	1	27.48		4.27	1		1/5F		Ran&FinWll,(TkRec)	480 m / 525 y
2	120418	2751412	1647401	1	27.60		4.15	2	1111	5/4		qaw, midtorls, (trackrecord)	480 m / 525 y
3	67655	4175158	2146615	1	27.63		4.34	3		5/4F	35.00	SoonLed	480 m / 525 y
4	211107	3806756	1994675	1	27.70		4.21	6		4/7F	32.25	qaw,wide,drewclear	480 m / 525 y
5	167836	3212881	1863989	1	27.71		4.30	1		2/7F		ep,ran&finwell	480 m / 525 y
6	95263	2720253	1489722	1	27.71		4.08	3	1111	7/4		quickaway, (trackrecord)	480 m / 525 y
7	38522	3033044	1717609	1	27.74		4.28	3		4/5F		SoonLed	480 m / 525 y
8	259670	3576671	1994675	1	27.74		4.24	6		1/1F		qaw,w,soonled	480 m / 525 y
9	284673	2589596	1535955	1	27.76		4.13	2	4221	4/6F		msdbrk, ep, (trackrecord)	480 m / 525 y

Looking at the race with this quickest time (race_id = 970811) at the following url http://www.greyhound-data.com/d?r=970811 We can see that the dog finished first in 29.12 however the dog that finished fourth finished in 20.58 which is clearly an error.



country full meeting		stadium	Monmore	race name	Monmore (Wolverhampton) 11 FEB 2002 HT 6
date	11 FEB 2002	race	6	going	0.10
distance	480 m / 525 y	type	flat_race	grade	A5
winner	0 £	second	0 £	third	0 £
trackrecord	27.95 sec		Larkhill Jo		7 JUL 1997
fastest of year	28.00 sec		Frisby Forte		23 AUG 2002
last q best	28.34 sec		Magna Mint		18 OCT 2001
this q best	28.51 sec		Goleen Melody		5 JAN 2002
q avg wintime	29.57 sec	avg time	29.87 sec		this q 547 dogs
comment					

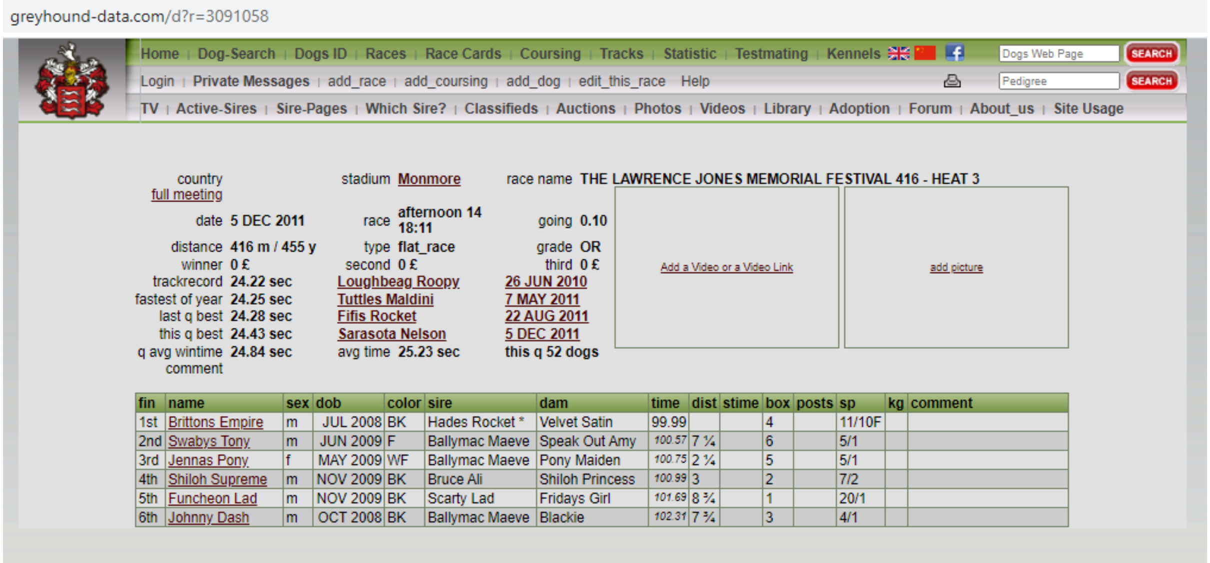
fin	name	sex	dob	color	sire	dam	time	dist	stime	box	posts	sp	kg	comment
1st	Supreme Agassi	m	JUN 2000	BK	Toms the Best	Westmead Josie	29.12		4.60	2	2111	5/2F		CrowdedStt and first, ran on
2nd	Hawkswood Nipper	m	MAR 2000	BD	Mustang Jack	Last Action	29.22		4.69	6	5642	4/1		CrowdedRunUp ran on
3rd	Floodhall Lad	m	SEP 1999	WBD	Smooth Rumble *	Lucky Number	29.46		4.69	5	6555	4/1		slowly away, crowded run up a
4th	Cushie Princess	f	SEP 1999	BEW	Vintage Prince	Cushie Amazing	20.58		4.58	3		5/1		quickly away, crowded first a
5th	Soviet Sea	m	AUG 1998	BK	Come On Ranger	Soviet Atlantic	29.59		4.57	4		7/2		early p crowded first and thir
6th	Butt Hatch Lad	m	AUG 1998	WBD	Lemon Rob	Highside	29.71		4.65	1		5/1		crowded first and second

Now lets have a look at the races with the slowest times

```
pd.read_sql_query('''
SELECT p.*, r.distance
FROM positions p
LEFT JOIN races r ON
    p.race_id = r.race_id
WHERE p.time IS NOT NULL and r.distance = '480 m / 525 y'
ORDER BY time desc
limit 10''', cnx)
```

	position_id	race_id	dog_id	fin	time	dist	stime	box	posts	sp	kg	comment	distance
0	299530	3095859	1609597	6	101.35	8 ½	None	1		6/1			480 m / 525 y
1	228998	3084516	1746667	6	101.12	4	None	3		7/4F			480 m / 525 y
2	317069	3105206	1600169	6	101.09	4 ½	None	1		10/1			480 m / 525 y
3	317103	3105211	1675189	6	100.93	3	None	5		10/1			480 m / 525 y
4	229030	3084521	1423132	6	100.91	HD	None	5		5/2			480 m / 525 y
5	229026	3084521	1617492	5	100.89	4	None	1		5/1			480 m / 525 y
6	229418	3079698	1804723	6	100.83	2	None	4		4/1			480 m / 525 y
7	228996	3084516	1675038	5	100.80	NK	None	1		5/1			480 m / 525 y
8	228999	3084516	1804242	4	100.77	2 ¾	None	4		9/2			480 m / 525 y
9	45448	1248731	958881	6	100.74	1 ¾	None	1		7/2			480 m / 525 y

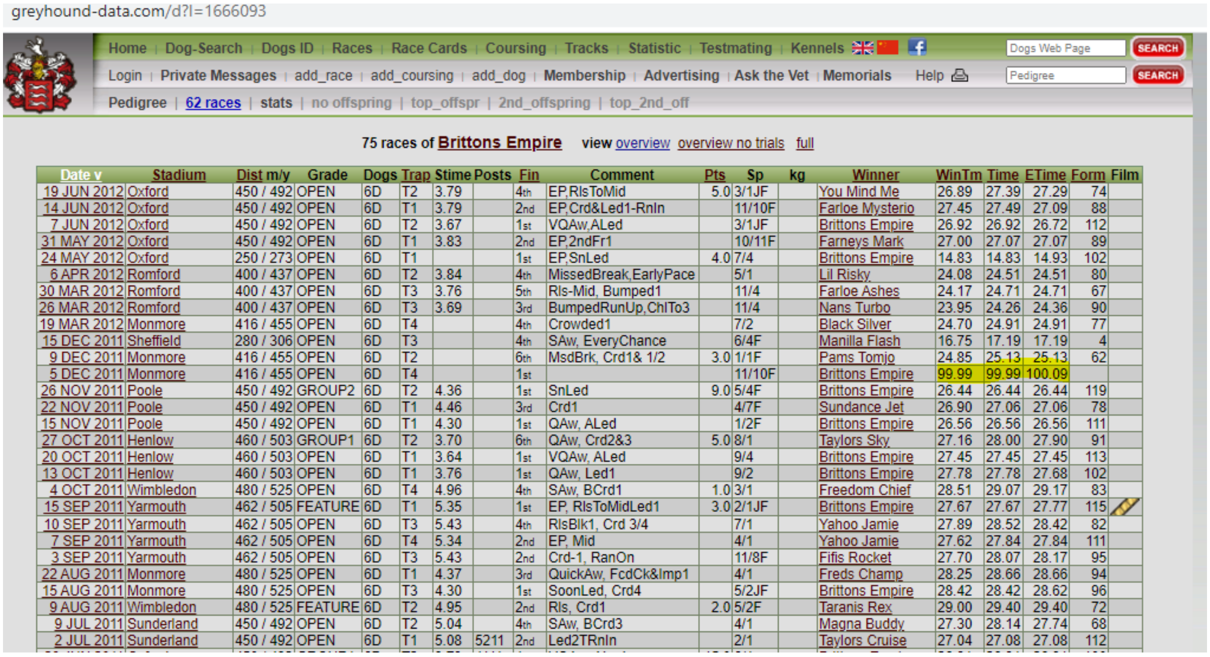
There looks to be something strange going on here. If we inspect the race with the slowest times 3091058 http://www.greyhound-data.com/d?r=3091058 we can see that the first dog recorded a time of 99.99 and the rest of the times are in italics. These times are unusually slow.


greyhound-data.com/d?r=3091058
Home | Dog-Search | Dogs ID | Races | Race Cards | Coursing | Tracks | Statistic | Testmating | Kennels
Dogs Web Page
SEARCH
Login | Private Messages | add_race | add_coursing | add_dog | edit_this_race Help
Pedigree
SEARCH
TV | Active-Sires | Sire-Pages | Which Sire? | Classifieds | Auctions | Photos | Videos | Library | Adoption | Forum | About_us | Site Usage
Add a Video or a Video Link
add picture

country full meeting		stadium	Monmore	race name	THE LAWRENCE JONES MEMORIAL FESTIVAL 416 - HEAT 3
date	5 DEC 2011	race	afternoon 14 18:11	going	0.10
distance	416 m / 455 y	type	flat_race	grade	OR
winner	0 £	second	0 £	third	0 £
trackrecord	24.22 sec	Loughbeag Roopy		26 JUN 2010	
fastest of year	24.25 sec	Tuttles Maldini		7 MAY 2011	
last q best	24.28 sec	Fifis Rocket		22 AUG 2011	
this q best	24.43 sec	Sarasota Nelson		5 DEC 2011	
q avg wintime	24.84 sec	avg time	25.23 sec	this q 52 dogs	
comment					

fin	name	sex	dob	color	sire	dam	time	dist	stime	box	posts	sp	kg	comment
1st	Brittons Empire	m	JUL 2008	BK	Hades Rocket *	Velvet Satin	99.99			4		11/10F		
2nd	Swabys Tony	m	JUN 2009	F	Ballymac Maeve	Speak Out Amy	*100.57*	7 ¼		6		5/1		
3rd	Jennas Pony	f	MAY 2009	WF	Ballymac Maeve	Pony Maiden	*100.75*	2 ¼		5		5/1		
4th	Shiloh Supreme	m	NOV 2009	BK	Bruce Ali	Shiloh Princess	*100.99*	3		2		7/2		
5th	Funcheon Lad	m	NOV 2009	BK	Scarty Lad	Fridays Girl	*101.69*	8 ¾		1		20/1		
6th	Johnny Dash	m	OCT 2008	BK	Ballymac Maeve	Blackie	*102.31*	7 ¾		3		4/1		

Clicking on the dog "Brittons Empire" that won the race we can see the history of its past races. This dog typically completes the race in under 30 seconds, and the 99.99 must be a data quality issue



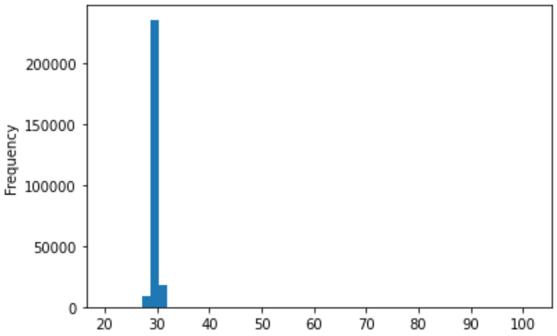
greyhound-data.com/d?l=1666093

Home | Dog-Search | Dogs ID | Races | Race Cards | Coursing | Tracks | Statistic | Testmating | Kennels
Dogs Web Page
SEARCH
Login | Private Messages | add_race | add_coursing | add_dog | Membership | Advertising | Ask the Vet | Memorials Help
Pedigree
SEARCH
Pedigree | 62 races | stats | no offspring | top_offspr | 2nd_offspring | top_2nd_off

75 races of **Brittons Empire** view overview overview no trials full

Date v	Stadium	Dist m/y	Grade	Dogs	Trap	Stime	Posts	Fin	Comment	Pts	Sp	kg	Winner	WinTm	Time	ETime	Form	Film
19 JUN 2012	Oxford	450 / 492	OPEN	6D	T2	3.79		4th	EP,RlsToMid	5.0	3/1JF		You Mind Me	26.89	27.39	27.29	74	
14 JUN 2012	Oxford	450 / 492	OPEN	6D	T1	3.79		2nd	EP,Crd&Led1-RnIn		11/10F		Farloe Mysterio	27.45	27.49	27.09	88	
7 JUN 2012	Oxford	450 / 492	OPEN	6D	T2	3.67		1st	VQAw,ALed		3/1JF		Brittons Empire	26.92	26.92	26.72	112	
31 MAY 2012	Oxford	450 / 492	OPEN	6D	T1	3.83		2nd	EP,2ndFr1		10/11F		Farneys Mark	27.00	27.07	27.07	89	
24 MAY 2012	Oxford	250 / 273	OPEN	6D	T1			1st	EP,SnLed	4.0	7/4		Brittons Empire	14.83	14.83	14.93	102	
6 APR 2012	Romford	400 / 437	OPEN	6D	T2	3.84		4th	MissedBreak,EarlyPace		5/1		Lil Risky	24.08	24.51	24.51	80	
30 MAR 2012	Romford	400 / 437	OPEN	6D	T3	3.76		5th	Rls-Mid, Bumped1		11/4		Farloe Ashes	24.17	24.71	24.71	67	
26 MAR 2012	Romford	400 / 437	OPEN	6D	T3	3.69		3rd	BumpedRunUp,ChlTo3		11/4		Nans Turbo	23.95	24.26	24.36	90	
19 MAR 2012	Monmore	416 / 455	OPEN	6D	T4			4th	Crowded1		7/2		Black Silver	24.70	24.91	24.91	77	
15 DEC 2011	Sheffield	280 / 306	OPEN	6D	T3			4th	SAw, EveryChance		6/4F		Manilla Flash	16.75	17.19	17.19	4	
9 DEC 2011	Monmore	416 / 455	OPEN	6D	T2			6th	MsdBrk, Crd1& 1/2	3.0	1/1F		Pams Tomjo	24.85	25.13	25.13	62	
5 DEC 2011	Monmore	416 / 455	OPEN	6D	T4			1st			11/10F		Brittons Empire	99.99	99.99	100.09		
26 NOV 2011	Poole	450 / 492	GROUP2	6D	T2	4.36		1st	SnLed	9.0	5/4F		Brittons Empire	26.44	26.44	26.44	119	
22 NOV 2011	Poole	450 / 492	OPEN	6D	T1	4.46		3rd	Crd1		4/7F		Sundance Jet	26.90	27.06	27.06	78	
15 NOV 2011	Poole	450 / 492	OPEN	6D	T1	4.30		1st	QAw, ALed		1/2F		Brittons Empire	26.56	26.56	26.56	111	
27 OCT 2011	Henlow	460 / 503	GROUP1	6D	T2	3.70		6th	QAw, Crd2&3	5.0	8/1		Taylors Sky	27.16	28.00	27.90	91	
20 OCT 2011	Henlow	460 / 503	OPEN	6D	T1	3.64		1st	VQAw, ALed		9/4		Brittons Empire	27.45	27.45	27.45	113	
13 OCT 2011	Henlow	460 / 503	OPEN	6D	T1	3.76		1st	QAw, Led1		9/2		Brittons Empire	27.78	27.78	27.68	102	
4 OCT 2011	Wimbledon	480 / 525	OPEN	6D	T4	4.96		4th	SAw, BCrd1	1.0	3/1		Freedom Chief	28.51	29.07	29.17	83	
15 SEP 2011	Yarmouth	462 / 505	FEATURE	6D	T1	5.35		1st	EP, RlsToMidLed1	3.0	2/1JF		Brittons Empire	27.67	27.67	27.77	115	
10 SEP 2011	Yarmouth	462 / 505	OPEN	6D	T3	5.43		4th	RlsBlk1, Crd 3/4		7/1		Yahoo Jamie	27.89	28.52	28.42	82	
7 SEP 2011	Yarmouth	462 / 505	OPEN	6D	T4	5.34		2nd	EP, Mid		4/1		Yahoo Jamie	27.62	27.84	27.84	111	
3 SEP 2011	Yarmouth	462 / 505	OPEN	6D	T3	5.43		2nd	Crd-1, RanOn		11/8F		Fifis Rocket	27.70	28.07	28.17	95	
22 AUG 2011	Monmore	480 / 525	OPEN	6D	T1	4.37		3rd	QuickAw, FcdCk&Imp1		4/1		Freds Champ	28.25	28.66	28.66	94	
15 AUG 2011	Monmore	480 / 525	OPEN	6D	T3	4.30		1st	SoonLed, Crd4		5/2JF		Brittons Empire	28.42	28.42	28.62	96	
9 AUG 2011	Wimbledon	480 / 525	FEATURE	6D	T2	4.95		2nd	Rls, Crd1	2.0	5/2F		Taranis Rex	29.00	29.40	29.40	72	
9 JUL 2011	Sunderland	450 / 492	OPEN	6D	T2	5.04		4th	SAw, BCrd3		4/1		Magna Buddy	27.30	28.14	27.74	68	
2 JUL 2011	Sunderland	450 / 492	OPEN	6D	T1	5.08	5211	2nd	Led2TRnIn		2/1		Taylors Cruise	27.04	27.08	27.08	112	

A histogram of race times and summary statistics confirms that those race times of 99+ must be an error.



	time	stime
count	262384.000000	260999.000000
mean	29.571383	4.497702
std	2.227873	1.138436
min	20.580000	0.400000
0.01%	27.840000	3.210000
0.1%	28.100000	4.100000
1%	28.430000	4.190000
25%	29.160000	4.380000
50%	29.480000	4.470000
75%	29.820000	4.550000
99%	30.830000	4.770000
99.9%	36.880710	5.150020
99.99%	100.565234	46.180040
max	101.350000	94.430000

we decide to remove any race from the data the contains times outside the following ranges

- 26 < time < 40
- 3 < stime < 6

Methodology

Ours is a supervised learning task with structured tabular data, therefore a natural choice is to use gradient boosting.

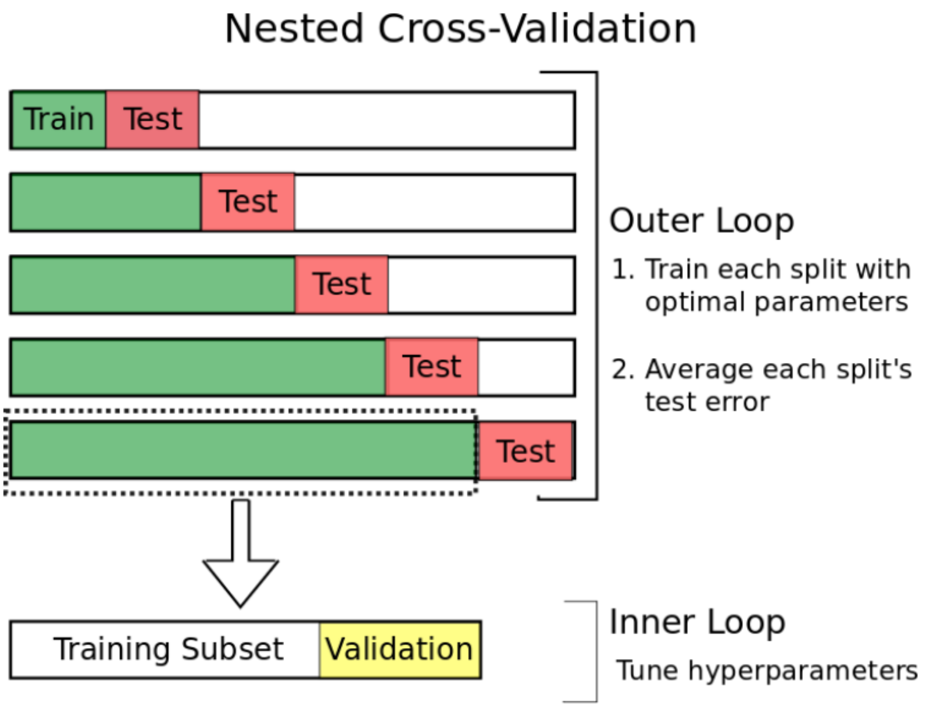
Features

The following features were constructed for each dog in the race. Summary statistics are calculated based on the last 25 days.

- Minimum time
- Average time
- Minimum stime (stime is time to the first bend)
- Average stime
- Average finish position
- Win percentage
- Place percentage (a place is coming in the top 2)
- Show percentage (a show is coming in the top 3)

Model Assessment

I decided to keep 2019 as a final test set, and used 2018 as a validation sample to test various modelling strategies (different features/varying amounts of training data etc). Since there is a time component to our problem I selected to use nested cross validation as follows



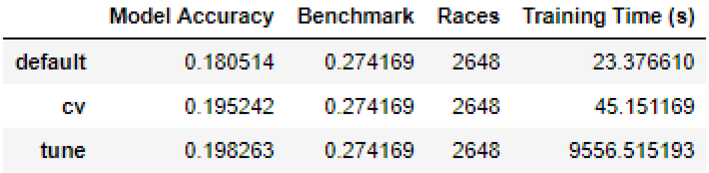
Source : https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

Training the Algorithm

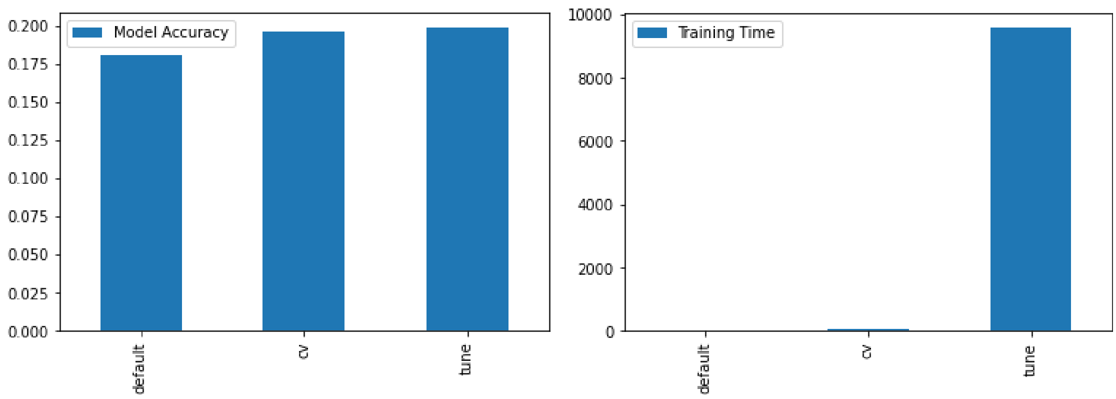
I have implemented three methods of training the model

1) **Default** - using the default parameters with no validation/early stopping
2) **CV** - Using the default parameters but tuning the number of boosting iterations using cross validation/early stopping to maximise accuracy.
3) **Tune** Using cross validation/early stopping to find the optimal hyperparameters and boosting iterations using Bayesian optimisation to maximise accuracy.

The default method produced the quickest result but also the lowest accuracy. CV took slightly longer but provided a noticeable jump in accuracy. Tuning the hyperparameters provided a modest improvement in accuracy for a very large increase in training time. It is noted that model accuracy is somewhat behind the benchmark, although it is ahead of random guessing (which is 0.167).



	Model Accuracy	Benchmark	Races	Training Time (s)
default	0.180514	0.274169	2648	23.376610
cv	0.195242	0.274169	2648	45.151169
tune	0.198263	0.274169	2648	9556.515193

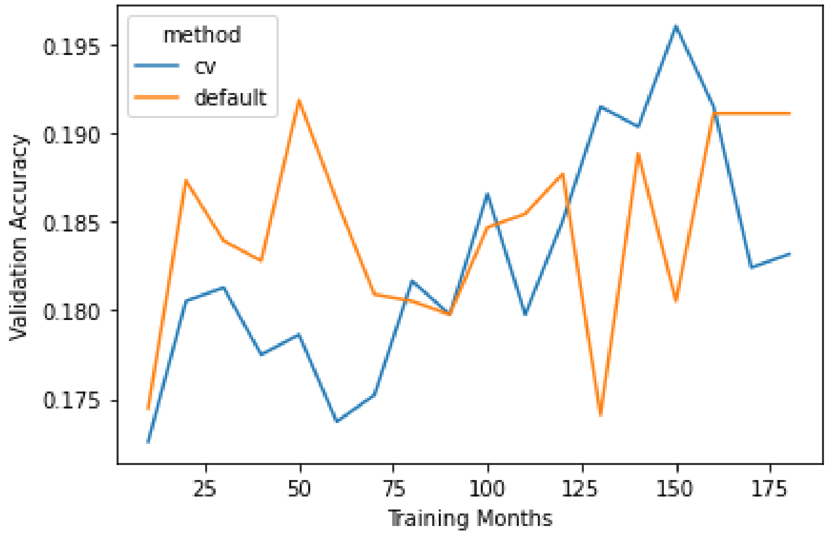


Complications

For a while I was getting terrible performance, worse than random change. I discovered this was because the algorithm requires the labels to be integers starting from 0, where as I had supplied them starting from 1! Subtracting one from both the winning box and benchmark resolved this.

Refinement

Learning curves were developed to understand the impact of providing the algorithm with more data. Two method of training the models (default, cv) were provided with increasing amounts of training data and the average accuracy across the 12 holdout months of 2018 are calculated.

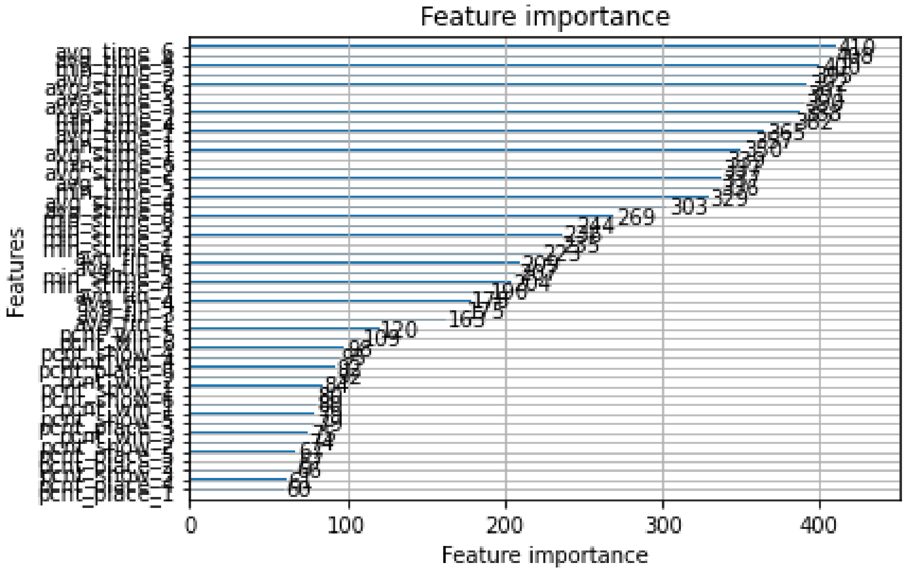


It appears that the cross validation method benefits more from receiving more data than the default parameters. This is most likely due to the fact that the default method trains for a fixed number of iterations, where as the cross validation method will keep training until the validation loss stops decreasing. It is also noted that the learning curve are quite noisy. This is due to the nature of the accuracy metric which is binary in nature. A small change in the predicted probabilities can suddenly change the predicted classification especially in a multi-class setting.

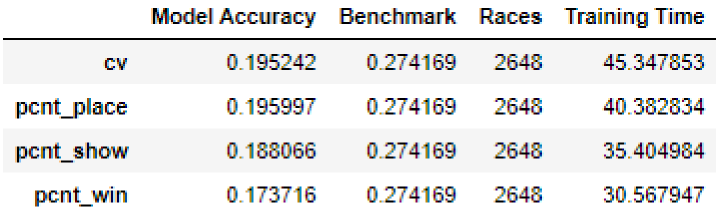
In also subsequent investigations a training method of CV and training months equal to 150 were used.

Feature Importance

For each of the 12 model runs used in the model assessment we produce a feature importance plot, an example of which is shown below. Average time appears at the top and the percent variables appear at the bottom.



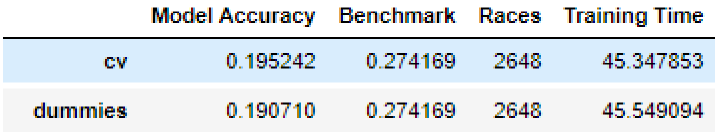
c. The results of which are as follows



	Model Accuracy	Benchmark	Races	Training Time
cv	0.195242	0.274169	2648	45.347853
pcnt_place	0.195997	0.274169	2648	40.382834
pcnt_show	0.188066	0.274169	2648	35.404984
pcnt_win	0.173716	0.274169	2648	30.567947

We can see that dropping pcnt_place has a marginal improvement relative to baseline, but taking out pcnt_show and pcnt_win has a detrimental effect on performance.

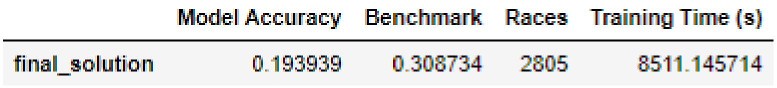
Lastly I constructued some dummy variables that indicated which of the dogs had the fastest average and minimum times. Unfortunately this reduced performance relative to the baseline



	Model Accuracy	Benchmark	Races	Training Time
cv	0.195242	0.274169	2648	45.347853
dummies	0.190710	0.274169	2648	45.549094

Results

To provide an honest assessment of model performance we evaluated the final model parameters/process on the previously unseen 2019 data.



	Model Accuracy	Benchmark	Races	Training Time (s)
final_solution	0.193939	0.308734	2805	8511.145714

Although it is pleasing that our model did not degrade in the test set and it is still ahead of random chance (0.167) we are way off the benchmark model.

Potential Improvements

Framing the problem as a regression problem where we try and predict the finish position for each dog would potentially make better use of the available information. In addition external data such as weather could be added, and other features based on a dogs past grade could be investigated further.