

Introducción a la plataforma Kaggle

Kaggle

- Introducción. Jose A. Guerrero.
- Marchamo “de facto” para data science (primeros = TRABAJO).
- Mas de 500.000 usuarios en todo el mundo (creciendo).
- Zona de test para los algoritmos mas avazandos (xgboost).
- [What has Kaggle learned from 2 million machine learning models?](#)
- [Lessons Learned from Running Hundreds of Kaggle Competitions](#)

Netflix


- El concurso del millón de dólares.
- Mejorar el algoritmo de recomendación de las películas.
- Se desarrolló en distintas fases y se obligaba a publicar al final de cada fase.
- Colaboración en los foros.
- Ensamblado de soluciones.
- 1111111111 -> 1111111000 y 0001111111
- No se llegó a implementar.
- <http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html>
- http://www.research.att.com/articles/featured_stories/2010_01/2010_02_netflix_article.html

Perfil Kaggle

smota




Verified account

MASTER  ?

Highest† **553rd** | Current† **575th**
/564,807

12,018.3 points
Joined 3 years ago
†Ranking method changed 13 May 2015 (?)



- Profile
- Results
- Scripts
- Forum
- Account
- Activity

Edit Profile

 10th / 345	 77th / 2619	 88th / 2226	 242nd / 2926	 499th / 5123	 104th / 634	 204th / 1076	 219th / 974	 18 Competitions
---------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------

<https://www.kaggle.com/santiagomota>

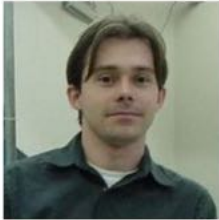









Santiago Mota (@mota_santiago)

Ranking

Kaggle Rankings

Kaggle users are allocated points for their performance in competitions. This page shows the current global ranking. For more information on how we calculate points, please visit the [user ranking wiki page](#).

🔍 Search for users


1st 187,877 pts  Gilberto Titericz 57 competitions Curitiba Brazil	2nd 171,495 pts  Μαριος Μιχαηλιδης 66 competitions Volos Greece	3rd 168,131 pts  Owen 42 competitions NYC United States	4th 158,518 pts  Stanislav Semenov 27 competitions Moscow Russian Federation	5th 147,687 pts  Alexander Guschin 20 competitions Moscow Russia
6th 136,923 pts  Abhishek 93 competitions Berlin Germany	7th 124,587 pts  Kohei Ozaki 63 competitions Tokyo Japan	8th 124,181 pts  Leustagos 42 competitions Belo Horizonte Brazil	9th 117,259 pts  Gert 23 competitions Goes The Netherlands	10th 101,679 pts  Dmitry Efimov 34 competitions Moscow Russian Federation

<https://www.kaggle.com/users>


Datasets

Welcome to Kaggle Datasets


The best place to discover and seamlessly analyze publicly available data.

 **Dig In**

Explore a dataset with our in-browser analytics tool, Kaggle Scripts. You can also download it in an easy to read format.

 **Build**

Create your data science portfolio. Publish insights and code with Kaggle Scripts and it will be saved to your profile.

 **Connect**

Engage with other data scientists. Share feedback on other Kagglers' scripts, or ask a question in a dataset's forum.



k **Ocean Ship Logbooks (175...**
130 Scripts · 28 Topics



k **US Dept of Education: Colle...**
334 Scripts · 18 Topics



UCI **Iris**
176 Scripts · 6 Topics

<https://www.kaggle.com/datasets>

Puestos de trabajo

Data Science Jobs Board



Hiring?

Access 456056 data scientists.

Kaggle is the world's largest community of data scientists, statisticians, and machine learning engineers. Kagglers demonstrate the skills to solve the toughest problems across many industries.

Create a Job Listing



Seeking?

Browse top data science careers.

The jobs board sources career openings for data professionals like you. Subscribe to be notified of new opportunities in data science, machine learning, statistics, and other analytics jobs.

Search our listings

Unsubscribe

Follow @KaggleCareers

Featured Posts



★ Senior Developer / Development Expert for Machine Learning

SAP · Germany / Israel / Singapore
posted 22 hours ago

291 views

<https://www.kaggle.com/jobs>

Ranking. Points. Tiers.

- Tres niveles (va a cambiar): Novice, Kaggle, Master.
- Los puntos “decaen”.
- Se tienen en cuenta los votos, número de participantes, si se forma parte de un equipo.
- [Brainstorming](#) Cambios.

<https://www.kaggle.com/wiki/UserRankingAndTierSystem>

Puntos

Points

Kaggle users are allocated points for their performance in competitions. The overall user rankings are shown at <https://www.kaggle.com/users>. These are recalculated at the end of every competition, once results have been finalized.

More points are earned for better results, with the maximum achievable points based on the number of total participants in the competitions, and a multiplier on the competition known as the "User Rank Multiplier". For certain competitions (e.g. Getting Started, or competitions with a public ground truth) the user rank multiplier of a competition is set to zero, meaning the competition will have no impact on users' points.

The current formula for each competition divides the points among the team members according to the square root, decays the points for lower finishes, adjusts for the number of teams that entered the competition, and decays the points as time elapses from the competition end. For each competition, the formula is:

$$\left[\frac{100000}{\sqrt{N_{\text{teammates}}}} \right] [\text{Rank}^{-0.75}] [\log_{10}(1 + \log_{10}(N_{\text{teams}}))] \left[e^{-t/500} \right]$$

Points are always calculated with time decay fixed at the time of the most recent competition deadline. Between competition deadlines points do not decay and ranks will not change.

<https://www.kaggle.com/wiki/UserRankingAndTierSystem>

Santiago Mota (@mota_santiago)

Masters

- Quedar entre los 10 primeros en una competición.
- Quedar en el primer 10% en otra.
- Da acceso a concursos especiales.
- [Unos 1000](#).



A screenshot of a Kaggle user profile card for Santiago Mota. The card has a light green background. At the top left, it says "MASTER" next to a cluster icon and a question mark icon. Below this, it shows "Highest+ 553rd" and "Current+ 575th / 564,807". At the bottom left, it says "12,018.3 points" and "Joined 3 years ago". At the bottom right, there is a small photo of a man with short grey hair. A small note at the bottom says "+Ranking method changed 13 May 2015 (?)".

MASTER	?
Highest+	Current+
553rd	575th
	/ 564,807
12,018.3 points	
Joined 3 years ago	
+Ranking method changed 13 May 2015 (?)	

<https://www.kaggle.com/wiki/UserRankingAndTierSystem>

Santiago Mota (@mota_santiago)

Datos

kaggle Host Competitions Datasets Scripts Jobs Community smota Logout



\$25,000 • 1,602 teams

Expedia Hotel Recommendations

Merger and 1st Submission Deadline

Fri 15 Apr 2016

Fri 10 Jun 2016 (14 days to go)

Dashboard

- Home
- Data
- Make a submission
- Information
 - Description
 - Evaluation
 - Rules
 - Prizes
 - Timeline
- Forum
- Scripts
 - New Script
 - New Notebook
- Leaderboard
- My Team
- My Submissions

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
sample_submission.csv	.gz (3.52 mb)
destinations.csv	.gz (16.18 mb)
test.csv	.gz (65.92 mb)
train.csv	.gz (511.16 mb)

Expedia has provided you logs of customer behavior. These include what customers searched for, how they interacted with search results (click/book), whether or not the search result was a travel package. **The data in this competition is a random selection from Expedia and is not representative of the overall statistics.**

Expedia is interested in predicting which hotel group a user is going to book. Expedia has in-house algorithms to form **hotel clusters**, where similar hotels for a search

<https://www.kaggle.com/c/expedia-hotel-recommendations/data>

Santiago Mota (@mota_santiago)

Public/private leaderboard I

	SalePrice	SquareFeet	Type	LotAcres	Beds	Baths
	\$88k	719	HOME	1.64	1	1
	\$164k	2017	APT		3	2
	\$72k	697	APT		1	1
	\$85k	948	HOME	1.02	2	3
	\$271k	3375	APT		3	4
	\$482k	3968	APT		4	4
	\$88k	790	APT		1	2
	\$128k	1341	HOME	0.66	3	3
	\$235k	2379	APT		3	3
	\$309k	2495	HOME	0.21	3	4
	\$163k	1356	APT		1	1
	\$375k	3361	HOME	1.64	3	4
	\$98k	1060	HOME	0.05	1	1
	???	582	HOME	0.61	1	1
	???	1640	APT		2	3
	???	3546	HOME	0.4	4	4
	???	903	APT		2	2
	???	1096	HOME	0.04	3	4
	???	1280	HOME	0.15	2	2
	???	1139	APT		1	1

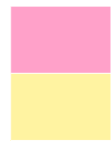
Training

Test

Predicted
\$41k
\$165k
\$280k
\$76k
\$128k
\$115k
\$94k

Submission

Public/private leaderboard II



Public Leaderboard \$14k
 Private Leaderboard \$15k

		SalePrice	SquareFeet	Type	LotAcres	Beds	Baths	
		\$88k	719	HOME	1.64	1	1	
		\$164k	2017	APT		3	2	
	MeanError	\$72k	697	APT		1	1	
		\$85k	948	HOME	1.02	2	3	
		\$271k	3375	APT		3	4	
		\$482k	3968	APT		4	4	
		\$88k	790	APT		1	2	
	Delta	Predicted	\$128k	1341	HOME	0.66	3	3
	-\$9k	\$41k	\$235k	2379	APT		3	3
	\$20k	\$165k	\$309k	2495	HOME	0.21	3	4
	-\$14k	\$380k	\$163k	1356	APT		1	1
	-\$6k	\$76k	\$375k	3361	HOME	1.64	3	4
	\$13k	\$128k	\$98k	1060	HOME	0.05	1	1
	-\$14k	\$115k	\$50k	582	HOME	0.61	1	1
	-\$12k	\$94k	\$145k	1640	APT		2	3
			\$394k	3546	HOME	0.4	4	4
			\$82k	903	APT		2	2
			\$105k	1096	HOME	0.04	3	4
			\$129k	1280	HOME	0.15	2	2
			\$106k	1139	APT		1	1

Training

Test

Submission

Leaderboard



\$25,000 • 1,602 teams

Expedia Hotel Recommendations

Fri 15 Apr 2016

Merger and 1st Submission Deadline

Fri 10 Jun 2016 (14 days to go)

Dashboard

Public Leaderboard - Expedia Hotel Recommendations

This leaderboard is calculated on approximately 33% of the test data. The final results will be based on the other 67%, so the final standings may be different.

See someone using multiple accounts?
[Let us know.](#)

#	Δ1w	Team Name	* in the money	Score	Entries	Last Submission UTC (Best - Last Submission)
1	↑3	Victor	*	0.51439	22	Thu, 26 May 2016 12:39:44
2	↑1	Viper	*	0.51402	14	Wed, 25 May 2016 18:37:51
3	↓1	AG100	👤 *	0.51306	22	Fri, 27 May 2016 00:46:52

<https://www.kaggle.com/c/expedia-hotel-recommendations/leaderboard>

Santiago Mota (@mota_santiago)

Forum



\$25,000 • 1,604 teams

Expedia Hotel Recommendations

Fri 15 Apr 2016

Merger and 1st Submission Deadline

Fri 10 Jun 2016 (14 days to go)

Dashboard

Competition Forum

New topic

Stop Watching

Search

1 2 3 4 5 6 7 8 9

Votes	166 topics, 1,076 posts		Replies	Views	Last Post
62	Data leak by Adam, 34 days ago	📌	46	12403	YongXien Chng yesterday
37	Welcome! by Adam, 41 days ago	📌	38	4738	Adam 6 days ago
1	 Interactive booking trends by Andrey Vykhodtsev, 18 days ago		1	0	Kendo 2 hours ago

<https://www.kaggle.com/c/expedia-hotel-recommendations/forums>

Santiago Mota (@mota_santiago)

Scripts

Highest Votes ▾ All Languages ▾ All Output Types ▾ All Competitions ▾ My Scripts







Feedback



Explore

Run one-click analyses, no local environment or data download needed



• • •

-  **Exploratory Analysis Rossmann**
last run 2 months ago by [thie1e](#) in [Rossmann Store Sales](#)
[33 comments](#) · [52 forks](#) · [30530 views](#) · [RMarkdown](#) ·  198 ↑
-  **Digging into Springleaf data**
last run 5 months ago by [Darragh](#) in [Springleaf Marketing Response](#)
[28 comments](#) · [58 forks](#) · [27238 views](#) · [RMarkdown](#) ·  169 ↑
-  **0.2748 with RF and log transformation**
last run 7 months ago by [arnaud demytt](#) in [Caterpillar Tube Pricing](#)
[31 comments](#) · [116 forks](#) · [12222 views](#) · [R](#) ·  146 ↑

118	new	 Vincent.Y	0.67306	10	Fri, 29 Jan 2016 02:41:01 (-23.4h)
119	+81	 apshreyans	0.67306	55	Fri, 29 Jan 2016 10:30:47 (-30.5h)

<https://www.kaggle.com/scripts?sortBy=votes>

Al empezar el concurso

- Tipo de concurso. ¿Alguno anterior?
- Cantidad de datos.
- Métrica de evaluación ([library\(Metrics\)](#)) y ([General](#)).
- Fechas límite.
- Partición public/private leaderboard.
- Suscribirse al foro.
- Buscar en Github.
- Leer las condiciones.
- Reproducción de la solución final.

<https://www.kaggle.com/wiki/WinningModelDocumentationTemplate>

Santiago Mota (@mota_santiago)

Estrategias

- Formación de equipos (límites).
- Scripts.
- Número de submissions al día.
- ¿Sobre cuantos modelos se hará la evaluación final?
- Elección de los modelos para el private leaderboard.
- Gestión de tiempos (dedicación).
- ¿Me fio del public leaderboard (overfitting)?
- Foro durante el concurso y al finalizar (huevos de pascua).

<https://www.kaggle.com/wiki/UserRankingAndTierSystem>

Varios

- Data leakage.
- Cuentas anónimas / imagen.
- Preguntar en el foro (puntos).
- Confirmación por SMS.
- Seed (xgboost).
- Titanic.
- Digit recognizer.
- 50% python, 40% R, 10% otros.
- Metodología de trabajo ([inversion](#))

¿Cuanto cuesta?

- En Kaggle el coste (premios incluidos) es de unos 100.000\$.
- Se incluye la preparación, seguimiento y análisis.
- Opciones gratuitas:
- Como profesor, en ese caso se limita y los alumnos y no se da asistencia (Kaggle Inclass).
- Con un proyecto que les parezca interesante.

Otras plataformas

- [CrowdAnalytics](#)
- [DrivenData](#)
- [Devpost](#)
- [Innocentive](#)
- [TunedIT](#)
- Enlaces a competiciones en [Kdnuggets](#)

Concursos presenciales

- Fin de semana vs extensos en el tiempo.
- Dotación económica.
- En equipo (casi siempre).
- Uso de otras “soft-skills”.
- Mas valor de la idea/presentación vs. datos/algoritmo.
- Limitaciones: Tiempo, datos, presentación.

Gracias

Datos de contacto:

Santiago Mota Herce

Teléfono: 670702852

Twitter: @mota_santiago

E-mail: santiago_mota@yahoo.es

LinkedIn: <https://es.linkedin.com/in/santiagomota>