

Presentación

- Físico. MBA y Máster BI y BigData EOI (2012/2013)
- M&A. CEO empresa náutica I+D+I
- Consultor freelance
- Kaggle Máster
- Socio Asociación de Usuarios de R de España
- santiago_mota@yahoo.es
- <http://es.linkedin.com/in/santiagomota>

Capítulo 1

Kaggle

Kaggle

- Kaggle
- Marchamo “de facto” para data science (primeros = TRABAJO)
- Mas de 100.000 usuarios activos en todo el mundo (creciendo)
- Zona de test para los algoritmos mas avanzados (xgboost)
- What has Kaggle learned from 2M ML models?
- Lessons Learned from Running Hundreds of Kaggle Competitions


Netflix

- 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

Kaggle


- ¿Por qué puede ser interesante para vosotros? Conocimientos, puestos de trabajo, metodologías, algoritmos y contactos.
- ¿Qué aporta en referencia con el Open Data? Datos, código, traspaso de conocimientos, open data.
- Elegir el nombre con cuidado.
- Apuntarse a los foros.

Perfil





Santiago Mota
Senior Data Scientist at Freelance
Madrid, Spain
Joined 6 years ago · last seen in the past day
[G](#) [T](#) [in](#)


Followers 355
Following 1118



Competitions Master


[Home](#) [Competitions \(63\)](#) [Kernels \(17\)](#) [Discussion \(84\)](#) [Datasets](#) [...](#) [Edit Profile](#)


Competitions Master

Current Rank
346
of 102,751
Highest Rank
238


Kernels Contributor

Unranked

Discussion Contributor

Unranked



1


4



9


ECML/PKDD 15: Taxi Trip ...
4 years ago · Top 3%


10th
of 345



Allstate Claims Severity
2 years ago · Top 2%


37th
of 3055



Prudential Life Insurance A...
3 years ago · Top 3%

77th
of 2619



0


0



0


SMH Data Cleaning: Scale...
a year ago

3
votes


Kernel_EDA_R
2 years ago

1
vote


Log_Mean_Plus LB: 0.47 Ve...
3 years ago

1
vote


1


2


34


Weather at Berlin US Airport
3 years ago

14
votes


Weather at Berlin US Airport
3 years ago

8
votes


Congrats to new Masters
3 years ago


7
votes




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

Santiago Mota (santiago_mota@yahoo.es)



Perfil hace unos años

 Host Competitions Datasets Scripts Jobs Community ▾ smota Logout

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




10th/345

TOP 10%




77th/2619

TOP 10%




88th/2226

TOP 10%




242nd/2926

TOP 10%




499th/5123

TOP 25%




104th/634

TOP 25%



204th/1076

TOP 25%






































219th/974

18

Competitions

Rankings

Kaggle Rankings							
Competitions		Kernels	Discussion	Learn more about rankings ›			
 136 Grandmasters		 1,150 Masters		 4,071 Experts		 46,627 Contributors	
						 51,784 Novices	
Rank	Tier	User			Medals		Points
346		 Santiago Mota	joined 6 years ago		 1  4  9		17,106
1		 bestfitting	joined 3 years ago		 18  4  0		241,899
2		 Giba	joined 7 years ago		 44  32  25		173,211
3		 Μαριος Μιχαηλιδης KazAnova	joined 6 years ago		 33  36  28		138,564
4		 Pavel Pleskov	joined 4 years ago		 9  16  7		122,764
5		 ZFTurbo	joined 3 years ago		 15  19  9		115,965

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

Santiago Mota (santiago_mota@yahoo.es)



Ranking, points, tiers

- Cinco niveles: Novice (44.000), Contributor (44.000), Expert (3.800), Master (1.000) y Grandmaster (125).
- También para kernels y foro.
- Los puntos “decaen”.
- Son menos si se forma parte de un equipo.
- Medallas.

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
🥉 Bronze	Top 40%	Top 40%	Top 100	Top 10%
🥈 Silver	Top 20%	Top 20%	Top 50	Top 5%
🥇 Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

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

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:

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

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/progression>





Datasets

697 featured datasets

Sort by Most Votes

Featured All Mine Upvoted

Q Search

592		IMDB 5000 Movie Dataset 5000+ movie data scraped from IMDB website chuansun76 · updated 10 months ago	38,424 downloads 74 comments
565		European Soccer Database 25k+ matches, players & teams attributes for European Professional Football Hugo Mathien · updated 9 months ago	28,880 downloads 94 comments
542		Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine Andrea · updated 8 months ago	27,516 downloads 61 comments
468		Human Resources Analytics Why are our best and most experienced employees leaving prematurely? ludoben · updated 7 months ago	26,357 downloads 85 comments

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

Puestos de trabajo

Data Science Jobs Board



Hiring?

Access thousands of 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



★ Data Scientist Machine Learning

Booking.com · Tel Aviv
posted yesterday

229
views

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

Santiago Mota (santiago_mota@yahoo.es)



Wiki

Learning Data Science

- [What is Data Science?](#)
- [Data Science Tutorials](#)
- [Software for Data Scientists](#)
- [Statistical and Machine Learning Algorithms](#)
- [External Data Science Books, Courses and References](#)
- [Data Science Blogs](#)
- [Data Sources](#)

Information for Participants

- [Kaggle Member FAQ](#)
- [Entering your First Submission](#)
- [Forming a Team](#)
- [Leaderboard](#)
- [Competition Scoring Metrics](#)
- [Strategy and Tactics](#)
- [Solutions to past competitions](#)
- [User Ranking and Tiers](#)
- [Winning Model Documentation Template](#)
- [Kaggle Connect Program](#)
- [Open Competition Data Sets](#)

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

Overview

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Submit Predictions](#)

Overview

Description

Evaluation


Frequently Asked Questions

Tutorials

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description



Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills


- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Santiago Mota (santiago_mota@yahoo.es)



Datos










House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

1,875 teams · 3 years to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Submit Predictions](#)

Training Data

 sample_submission.cs...	<h3>data_description.txt</h3> 13.06 KB Download
 test.csv	
 train.csv	
 data_description.txt	
 sample_submission.cs...	
 test.csv.gz	
 train.csv.gz	

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

Santiago Mota (santiago_mota@yahoo.es)













Kernels

Overview Data **Kernels** Discussion Leaderboard Rules New Kernel

Sort by **Most Votes**

All Mine All Languages All Output Types

384			Regularized Linear Models run 3 months ago by Alexandru Papiu	Py	165
339			Comprehensive data exploration with Python run 6 days ago by Pedro Marcelino	Py	159
144			Detailed Exploratory Data Analysis using R run 6 months ago by AiO (+151 / -430 / -183)	R	64
120			Detailed Data Exploration in Python run 3 months ago by Angela	Py	45
116			Ensemble Model: Stacked Model Example run 10 months ago by JMT5802	R	59

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/kernels>

Santiago Mota (santiago_mota@yahoo.es)













Foro

Overview Data Kernels **Discussion** Leaderboard Rules [New Topic](#)

251 topics and kernels [\(Unsubscribe\)](#) Sort by **Most Votes**

All Mine | Upvoted Topics & Kernels Search

384			Regularized Linear Models last run 10 months ago by Alexandru Papiu	last comment by Jaswinder Singh 3 days ago	165
339			Comprehensive data exploration with Python last run 5 months ago by Pedro Marcelino	last comment by Tommy Jiang an hour ago	159
144			Detailed Exploratory Data Analysis using R last run 10 months ago by AiO	last comment by Pooja Babu 9 days ago	64
120			Detailed Data Exploration in Python last run 10 months ago by Angela	last comment by dataDuckling a month ago	45
116			Ensemble Model: Stacked Model Example last run 10 months ago by JMT5802	last comment by Andrew-Chang a month ago	59

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion>

Santiago Mota (santiago_mota@yahoo.es)



Rules

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)

[Submit Predictions](#)

One account per participant

You cannot sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts.

No private sharing outside teams

Privately sharing code or data outside of teams is not permitted. It's okay to share code if made available to all participants on the forums.

Team Mergers

Team mergers are allowed and can be performed by the team leader. In order to merge, the combined team must have a total submission count less than or equal to the maximum allowed as of the merge date. The maximum allowed is the number of submissions per day multiplied by the number of days the competition has been running.

Team Limits

There is no maximum team size.

Submission Limits

You may submit a maximum of 5 entries per day.

You may select up to 2 final submissions for judging.

Competition Timeline

Start Date: 8/30/2016 1:08 AM UTC

Merger Deadline: **None**

Entry Deadline: **None**

End Date: 3/1/2017 11:59 PM UTC

- Due to the public nature of the data, this competition does not count towards Kaggle ranking points.
- We ask that you respect the spirit of the competition and do not cheat. Hand-labeling is forbidden.

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/rules>

Santiago Mota (santiago_mota@yahoo.es)



Team

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)





Manage Team

Team Name


[Save Team Name](#)

This name will appear on your team's leaderboard position.

Team Members

		QualityExcellence	Leader
		smota (you)	Member

Invite Others

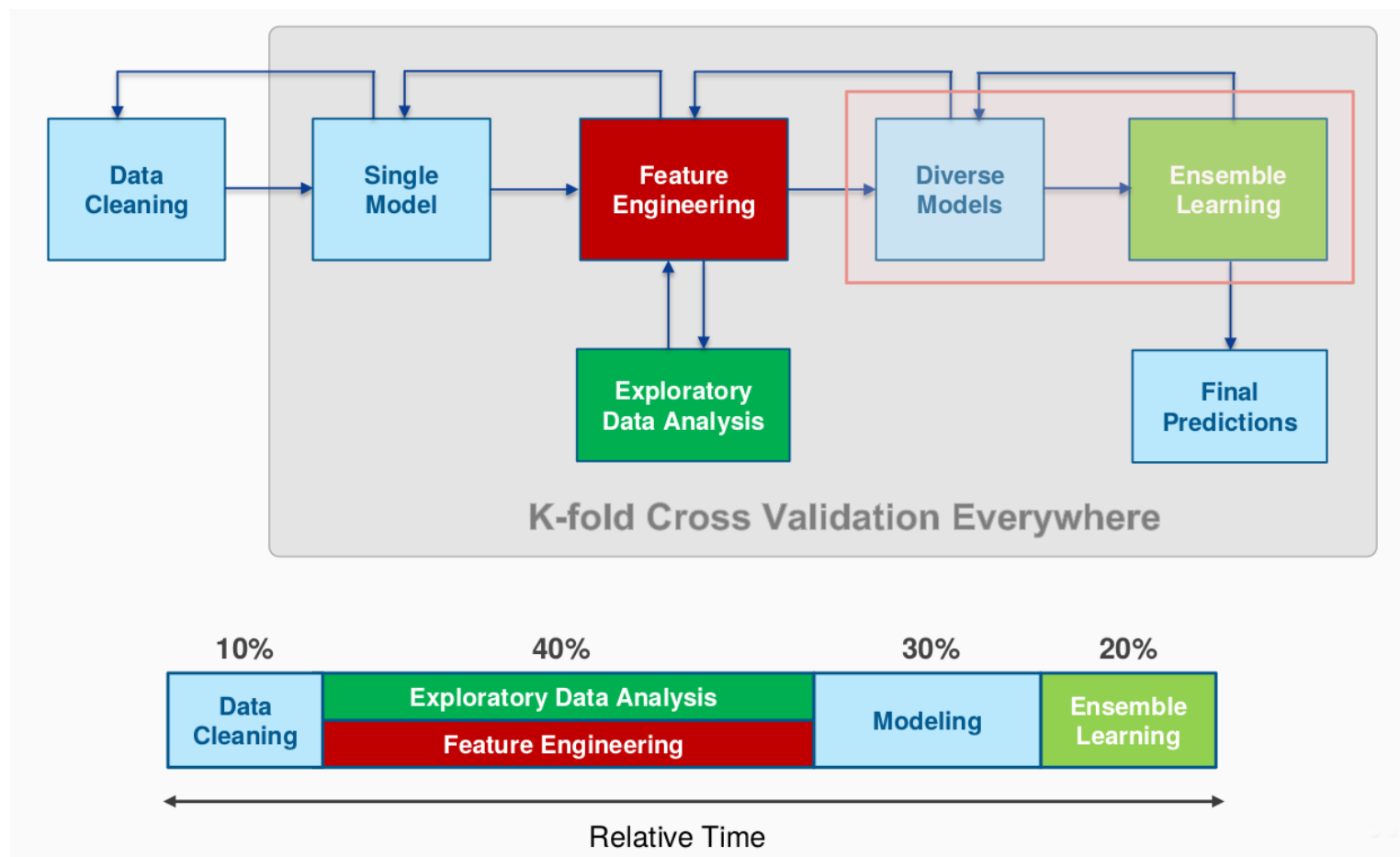
 **Merge with other Kaggle teams**

You are unable to merge with other teams on this competition.

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.

Proceso de trabajo



<https://www.slideshare.net/markpeng/general-tips-for-participating-kaggle-competitions>

Estrategias

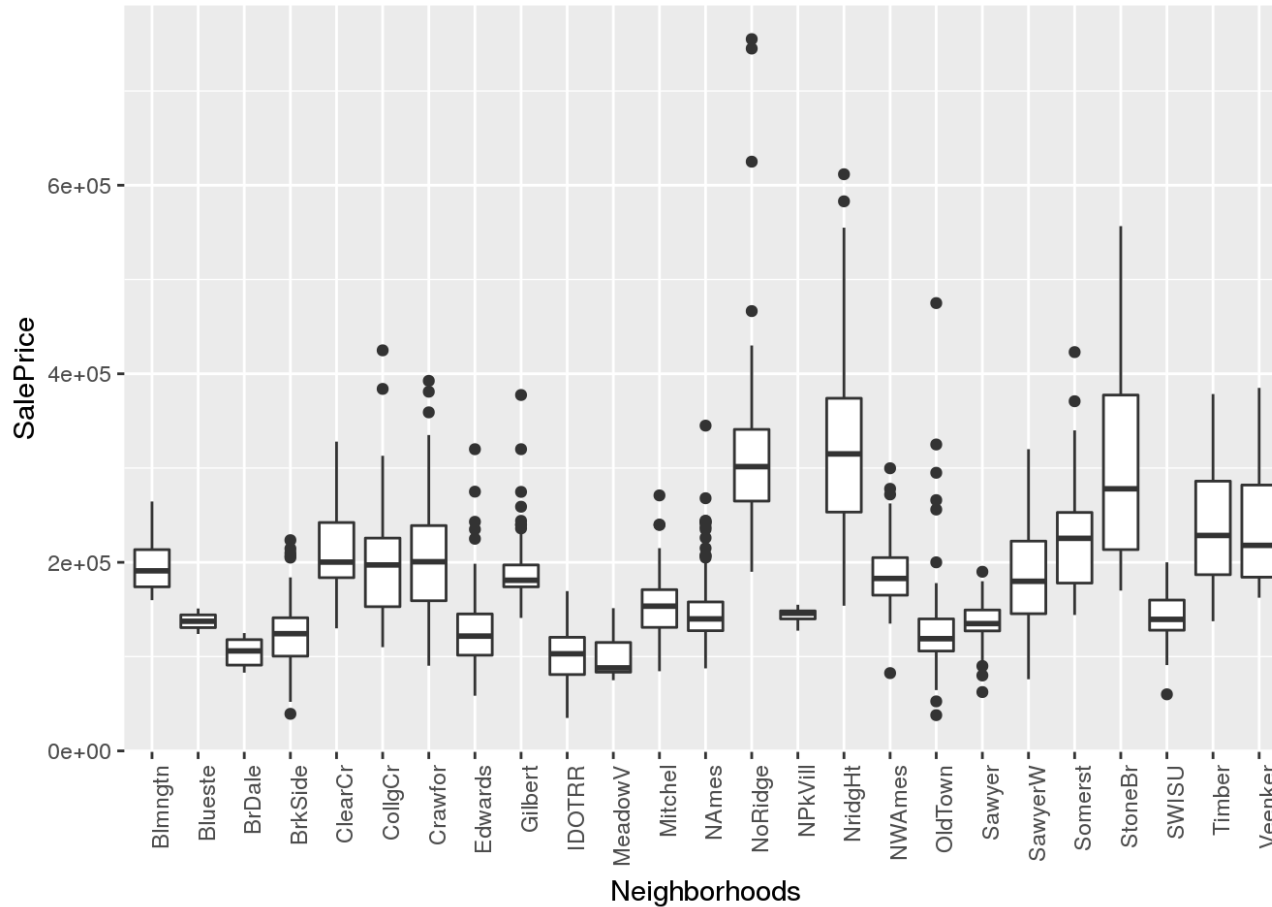
- Formación de equipos (límites).
- Kernels.
- Número de submissions al día.
- ¿Sobre cuantos modelos se hará la evaluación final?
- Elección de los modelos para el private leaderboard.
- Gestión de tiempos (dedicación).
- ¿Me fio del public leaderboard (overfitting)?
- Foro durante el concurso y al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

Varios

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

Concurso precios de alquileres

- [Página del concurso](#). [Tutoriales](#). [Kernels](#) y [Foro](#)



¿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 ([link](#)).
- DrivenData ([link](#)).
- Devpost ([link](#)).
- Innocentive ([link](#)).
- TunedIT ([link](#)).
- Enlaces a competiciones en Kdnuggets ([link](#)).

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>

Capítulo 2

Donde continuar

Bibliografía

- [Introducción a R](#) (castellano).
- [R para principiantes](#) (castellano).
- [An introduction to R](#) (inglés).
- [R programming for data science](#) (inglés).
- [R for data science](#) (inglés / pago).
- [R For Dummies](#) (inglés / pago).

Webs

- R bloggers
- KD nuggets
- Data Science Central
- GitHub
- Kaggle
- Medium

Chuletas

- Sobre R en general: [Una](#), [dos](#), [tres](#) y [cuatro](#)
- [Varias de Rstudio](#). Entre ellas Markdown, Rstudio o Shiny
- [ggplot2](#)
- [Expresiones regulares](#)
- [dplyr](#)
- [data.table](#)
- En Rstudio (\Help\Cheatsheets)

MOOC

- Introducción a R ([Datacamp](#)).
- Introduction to R ([Microsoft](#)).
- [R programming](#) Johns Hopkins. Peng y Leek ([data science](#)).
- [Statistical Learning](#). Stanford.

Otros

- Asociación de usuarios de R de España
- Meetup Grupo de Usuarios de R de Madrid
- Jornadas nacionales de usuarios de R
- Stackoverflow (poner [r] en la búsqueda)
- Glosario de Machine Learning de Google
- Essentials of Machine Learning Algorithms
- Pautas de Google en el estilo de programación
- BI and Big Data Open Sourced