



# Getting started with **kaggle**<sup>™</sup>



# Contents



01 | What is Kaggle?

02 | Skills Required

03 | Kaggle Features - Competitions, Kernels, Forums, Datasets, Tutorials

04 | Tips and Tricks

05 | Where to begin

# What is Kaggle




Online platform for Data Science

- Machine Learning Competitions
- Data Visualizations
- Code Sharing (Kernels)
- Datasets
- Discussion Forums (DS goldmine!!)
- Tutorials
- Blog

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


# Kaggle “Portfolios”



**Adarsh Chavakula**

India  
Joined 4 years ago · last seen in the past day


[in](#)

**Competitions Expert**

Followers 4  
Following 4


[Edit Profile](#)


[Home](#) [Competitions \(21\)](#) [Kernels \(22\)](#) [Discussion \(29\)](#) [Datasets \(2\)](#) ...


**Competitions Expert**

Current Rank  
**1235**  
of 73,787

Highest Rank  
**744**

**0**

**2**


**1**

[Rossmann Store Sales](#)  
🕒 · 2 years ago · Top 1%


**28<sup>th</sup>**  
of 3303


[Sberbank Russian Housing ...](#)  
🕒 · 7 months ago · Top 4%


**116<sup>th</sup>**  
of 3274

**Kernels Contributor**

**Unranked**

**0**

**0**


**1**

[How to cross validate prope...](#)  
🕒 · 7 months ago


**17**  
votes


[GLM - 0.61](#)  
2 years ago


**0**  
votes

**Discussion Contributor**

**Unranked**

**0**

**3**

**15**


[Did anyone see "This site h...](#)  
🕒 · 2 years ago

**9**  
votes

[Non XGBoost Success?](#)  
🕒 · 2 years ago


**6**  
votes

# Kaggle “Portfolios”



## Michael Jahrer

Graz, Austria  
Joined 8 years ago · last seen in the past day  
<http://www.operasolutions.com/>







**Competitions  
Grandmaster**

Followers 573


[Home](#) [Competitions \(58\)](#) [Kernels \(6\)](#) [Discussion \(161\)](#) [Followers \(573\)](#) [Contact User](#) [Follow User](#)

### Competitions Grandmaster






<b>Current Rank</b> <b>10</b> of 73,787	<b>Highest Rank</b> <b>6</b>	
 <b>13</b>	 <b>11</b>	 <b>2</b>
<b>Porto Seguro's Safe Driver ...</b> 🥇 · 2 months ago · Top 1%	<b>1<sup>st</sup></b> of 5169	
<b>Cervical Cancer Screening</b> 🥇 · 2 years ago · Top 3%	<b>1<sup>st</sup></b> of 40	


### Kernels Contributor






**Unranked**

 <b>0</b>	 <b>0</b>	 <b>0</b>
<b>XGB Feature Importance (P...</b> 2 years ago	<b>5</b> votes	
<b>naive XGB</b> 9 months ago	<b>2</b> votes	

### Discussion Expert



**Rank 9**  
of 45,781

 <b>12</b>	 <b>28</b>	 <b>73</b>
<b>1st place with representatio...</b> 🥇 · 2 months ago	<b>547</b> votes	
<b>Is 0.288 magical and optim...</b> 🥇 · 2 months ago	<b>60</b> votes	

# Skills Required



1

**Programming**  
Python or R

2

**Data Handling**

3

**Applied Machine Learning**

4

**Data Visualization**

Python - Seaborn/Matplotlib  
R - ggplot2

# Commonly used libraries



Task	Python	R
Data Handling	pandas	dplyr
Data Visualization	seaborn	ggplot2
Linear Models	scikit-learn	GLM
Gradient Boosting	XGBoost, LightGBM	XGBoost
Deep Learning	Tensorflow, Keras	-
Other ML algorithms	scikit-learn	caret
NLP	NLTK, spaCy	Tidyttext and others

# Competitions



- Conducted by companies, research organizations or non-profits.
- Generally run for a month or two (some may be significantly longer)
  - Limited number of submissions per day (usually 5/day)
- Different tiers
  - **Featured** - Most prestigious. Winners get \$\$, recruitment offers, bragging rights. All participants get points based on final standing.
  - **Research** - Generally more difficult. May not have prizes or points.
  - **Playground** - For beginners. Small Datasets. Lots of community support. No prizes or points but valuable learning



# Competitions <https://www.kaggle.com/competitions>

## 12 Active Competitions



### 2018 Data Science Bowl

Find the nuclei in divergent images to advance medical discovery

**Featured** · 3 months to go · 🧬 biology

**\$100,000**  
494 teams



### Statoil/C-CORE Iceberg Classifier Challenge

Ship or iceberg, can you decide from space?

**Featured** · 21 hours to go · 🌤️ weather, shipping, image data, binary classification

**\$50,000**  
2,511 teams



### Toxic Comment Classification Challenge

Identify and classify toxic online comments

**Featured** · 2 months to go · 💬 arguments, text data

**\$35,000**  
547 teams



### IEEE's Signal Processing Society - Camera Model Identification

Identify from which camera an image was taken

**Featured** · 17 days to go · 📷 image data

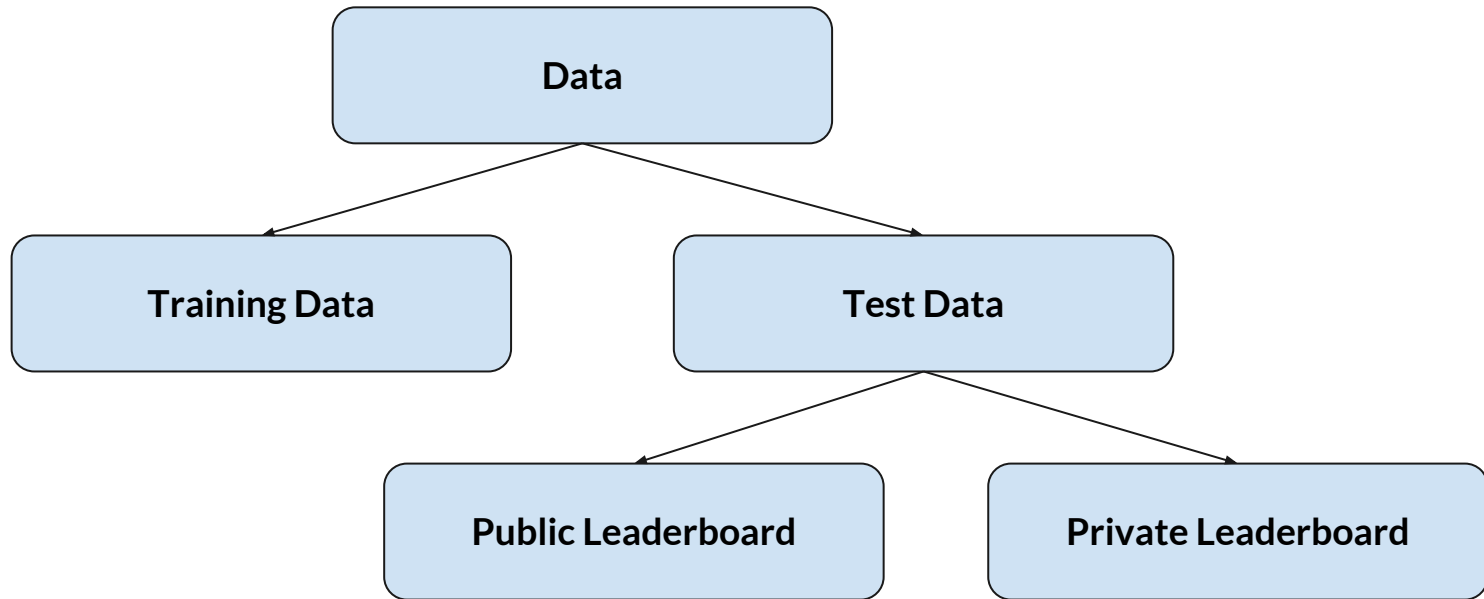
**\$25,000**  
354 teams

# Competition Types



- Different types of problems:
  - Regression, Classification, Recommendation, Optimization
- Different types of datasets depending on the problem
  - Tabular, Text (Natural Language Processing), Image (Computer Vision)
- Datasets are *usually* very well structured and clean. Very well defined objective.

# Competition Structure



# Kernels



- Public/Private codes and notebooks that can be run on the Kaggle server
- Kaggle's in-house code sharing mechanism
- Comes with a ton of pre-installed packages for Python, R and Julia.

<https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/kernels>

# Kernels

Select Kernel Type



## Script

```
import numpy as np # linear algebra
import pandas as pd # data processing,

# Input data files are available in the current working directory
# as well as in the "../input" directory

from subprocess import check_output
print(check_output(["ls", "../input"]))

# Any results you write to the current directory are saved as output files
```

- Python, R, RMarkdown
- Runs all the code, every time
- Ideal for fitting a model and competition submissions
- Shares code for review and RMarkdown reports

## Notebook

Introduction

```
# Loading in the training data
train = pd.read_csv("../train.csv")
```

- Jupyter Notebooks in Python or R
- Runs cells of code and Markdown
- Ideal for interactive data exploration and polished analysis
- Shares insights through code & commentary

# Discussion Forums



- Every competition has a public discussion forum
- Questions for the organizers, information about the datasets, idea sharing etc.
- Goldmine of knowledge!

<https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/discussion>

# Datasets



- Kaggle is home to several cool datasets!
- Can also publish your own datasets

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

# Tutorials



- Kaggle has a really useful tutorial section for beginners
- Has 4 learning tracks - Machine Learning, R, Data Visualization and Deep Learning

<https://www.kaggle.com/learn/overview>



# Blog - No Free Hunch <http://blog.kaggle.com/>

➤ The Official Blog of  
Kaggle.com

Q Search

## Categories

DATA SCIENCE NEWS (61)

KAGGLE NEWS (136)

KERNELS (41)

OPEN DATASETS (8)

TUTORIALS (49)

WINNERS' INTERVIEWS (218)

## Want to subscribe?

Email Address\*

## No Free Hunch

KAGGLE.COM

### Our Final Kaggle Dataset Publishing Awards Winners' Interviews (November 2017 and December 2017)

Megan Risdal | 01.24.2018

As we move into 2018, the monthly Datasets Publishing Awards has concluded. We're pleased to have recognized many publishers of high-quality, original, and impactful datasets. It was only a little over a year ago that we opened up our public Datasets platform to data enthusiasts all over the world to share their work. We've now reached almost 10,000 public datasets, making choosing winners each month a difficult task! These interviews feature the stories and backgrounds of the November and December ...

DATASET PUBLISHING AWARDS

### Reviewing 2017 and Previewing 2018

Anthony Goldbloom | 01.22.2018

2017 was a huge year for Kaggle. Aside from joining Google, it also marks the year that our community expanded from being primarily focused on machine learning competitions to a broader data science and machine learning platform. This year our public Datasets platform and Kaggle Kernels both grew ~3x, meaning we now also have a thriving data repository and code sharing environment. Each of those products are on track to pass competitions on most activity metrics in early 2018. To ...

### An Intuitive Introduction to Generative Adversarial Networks

Keshav Dhandhaniah | 01.18.2018



# **Tips and Tricks!**

# Competitions - Basics



- Start early!
- Track your progress and code versions - GitHub is the way to go
- Step 1 is always setting up Cross Validation strategy
- DO NOT TRUST THE PUBLIC LEADERBOARD
- Read everything on the discussion forum
- Draw inspiration from Kernels but don't rely on them too much

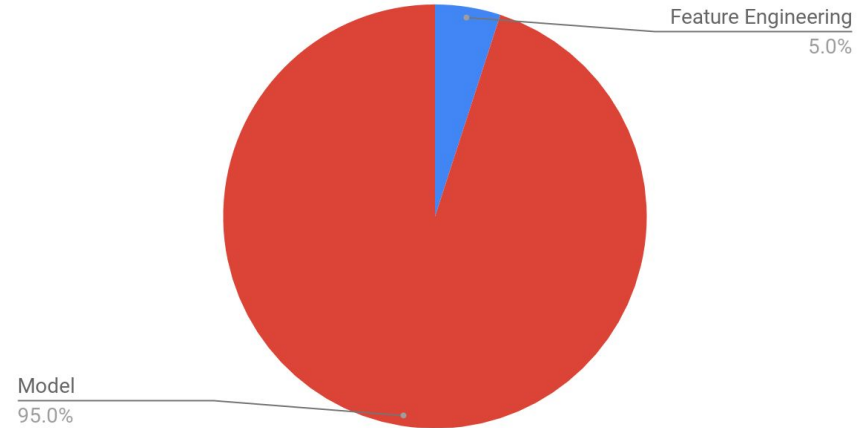
# Competitions - Feature Engineering



Contribution to final score



Time allocation of average participants



Feature Engineering is often the key differentiator between average participants and the top performers

# Competitions - Feature Engineering



- Normalize variables for deep learning, SVMs, linear models and other distance based algorithms
- Log transform variables which vary over several orders of magnitude
- One-hot encode categorical features
- Explore combinations of variables and their interactions
- Feature Selection - Identify and eliminate useless features

# Competitions - Modeling



- Explore all algorithms - Deep Learning does not always win
- Set up an effective cross validation strategy for model evaluation. Do not rely on the Leaderboard for evaluating model performance
- Model Ensembling is the key to a better score

<https://mlwave.com/kaggle-ensembling-guide/>



**Where do I start?**

# 1. I am an absolute beginner in Data Science



- Improve your programming proficiency by practising R/Python
- Take the Kaggle tutorials (<https://www.kaggle.com/learn/overview>)
- Try out the “*Playground*” competitions and work your way up



## 2. I can code but can't Data Science



- Take the Kaggle tutorials, especially ML and Data Viz (<https://www.kaggle.com/learn/overview>)
- Try out the “*Playground*” competitions and work your way up

### 3. I can code and know Machine Learning



- Master all commonly used applied machine learning libraries
- Dive into the “*Featured*” and “*Research*” competitions

# Parting Remarks



- Dive right in. Don't hesitate.
- Don't be discouraged by the difficulty or the leaderboard standings - focus on learning and applying new skills.



# Thank you

