# Kaggle walkthrough: Restaurant Sales Prediction


Bikash Agrawal
Co-Founder Boost AI
PhD Candidate
University of Stavanger

# Who are we?

Bikash Agrawal

PhD Candidate in big data and data processing

Lars Selsås

Experience from Silicon Valley and expert within machine learning

Henry V Iversen

Experience from start-ups and taking a master degree in Economics

Gilberto Titericz Jr

Number 1 in machine learning competition Kaggle

# Agenda

- Machine Learning
- Kaggle
- What is the competition goal?
- Why is this difficult?
- What data do we have?
- What can we learn from the data?
- Can machine learning help?
- What did I learn?

# Machine Learning

Machine Learning = computer learns patterns from data.

# Data Science Competition


NETFLIX
2006 - 2009

kaggle
Since 2010

CrowdANALYTIX

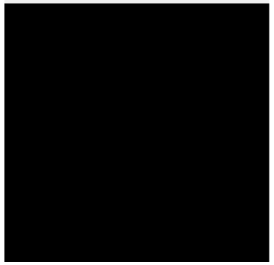
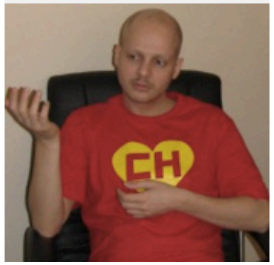
DRIVENDATA

# Kaggle

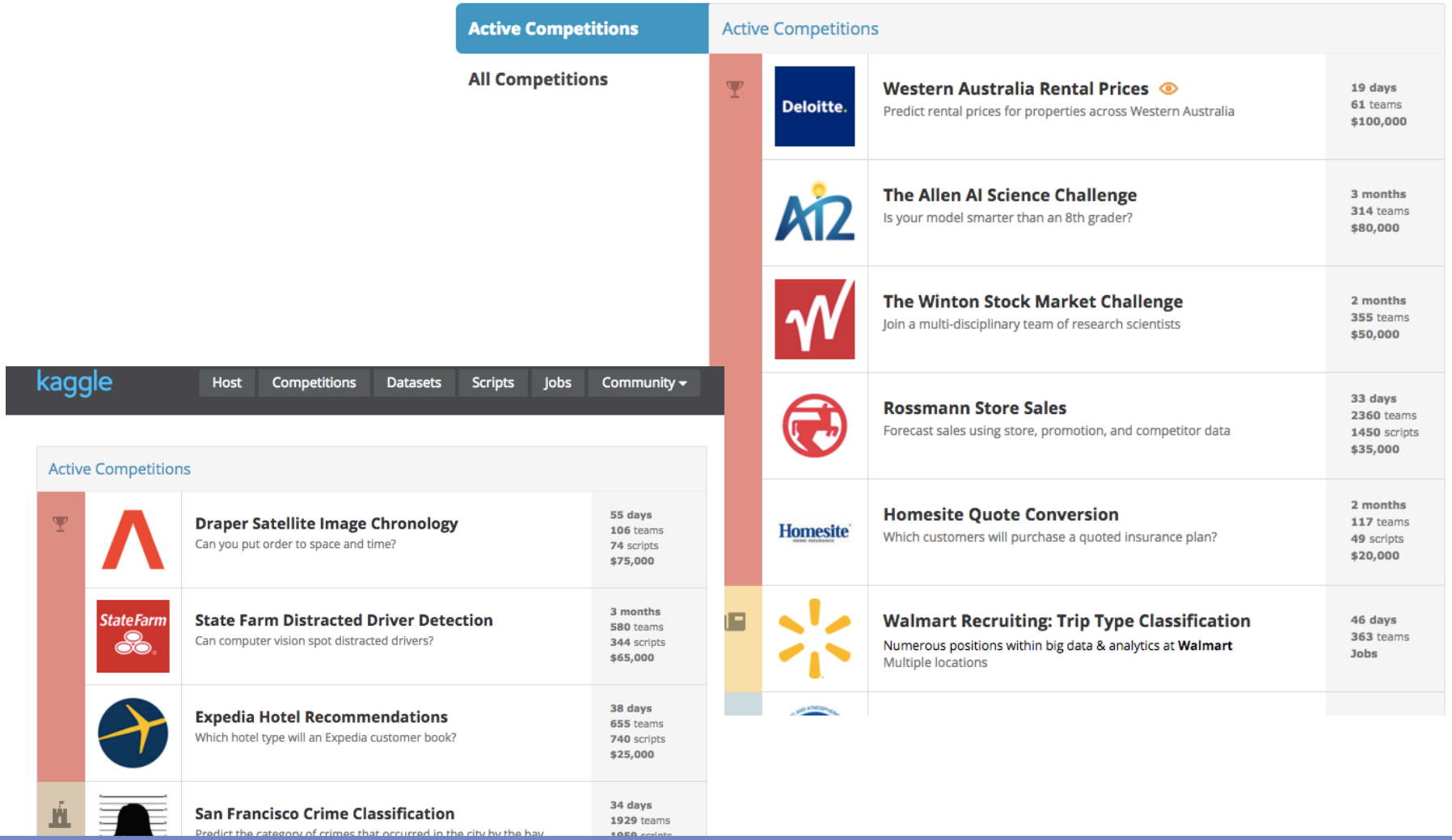
- Kaggle is the world's largest community of data scientists with over 300,000 registered users (www.kaggle.com).
- Kaggle was set up to demonstrate the power of team competitive forces for outperforming previous best benchmark model results.
- The Kaggle community of data scientists (Kagglers) comprises tens of thousands of PhDs from quantitative fields such as computer science, statistics, econometrics, maths and physics, and industries such as insurance, finance, science, and technology.
- Kagglers are also a diverse bunch, being spread across over 100 countries and 200 universities worldwide.

# Kaggle

| 1st | 2nd | 3rd | 4th | 5th |
|---|---|---|---|---|
| 208,807 pts | 203,405 pts | 184,404 pts | 149,538 pts | 140,435 pts |
|  |  |  |  |  |
| **Gilberto Titericz** | **Μαριος Μιχαηλιδης** | **Stanislav Semenov** | **Kohei** | **Owen** |
| 69 competitions<br>Curitiba<br>Brazil | 75 competitions<br>Volos<br>Greece | 34 competitions<br>Moscow<br>Russian Federation | 73 competitions<br>Tokyo<br>Japan | 42 competitions<br>NYC<br>United States |

| 6th | 7th | 8th | 9th | 10th |
|---|---|---|---|---|
| 126,558 pts | 121,763 pts | 119,023 pts | 117,812 pts | 114,566 pts |
|  |  |  |  |  |

# Why kaggle


Active Competitions
All Competitions
Active Competitions
Western Australia Rental Prices
Predict rental prices for properties across Western Australia
19 days
61 teams
$100,000
The Allen AI Science Challenge
Is your model smarter than an 8th grader?
3 months
314 teams
$80,000
The Winton Stock Market Challenge
Join a multi-disciplinary team of research scientists
2 months
355 teams
$50,000
Rossmann Store Sales
Forecast sales using store, promotion, and competitor data
33 days
2360 teams
1450 scripts
$35,000
Homesite Quote Conversion
Which customers will purchase a quoted insurance plan?
2 months
117 teams
49 scripts
$20,000
Walmart Recruiting: Trip Type Classification
Numerous positions within big data & analytics at Walmart
Multiple locations
46 days
363 teams
Jobs
kaggle
Host
Competitions
Datasets
Scripts
Jobs
Community
Active Competitions
Draper Satellite Image Chronology
Can you put order to space and time?
55 days
106 teams
74 scripts
$75,000
State Farm Distracted Driver Detection
Can computer vision spot distracted drivers?
3 months
580 teams
344 scripts
$65,000
Expedia Hotel Recommendations
Which hotel type will an Expedia customer book?
38 days
655 teams
740 scripts
$25,000
San Francisco Crime Classification
34 days
1929 teams

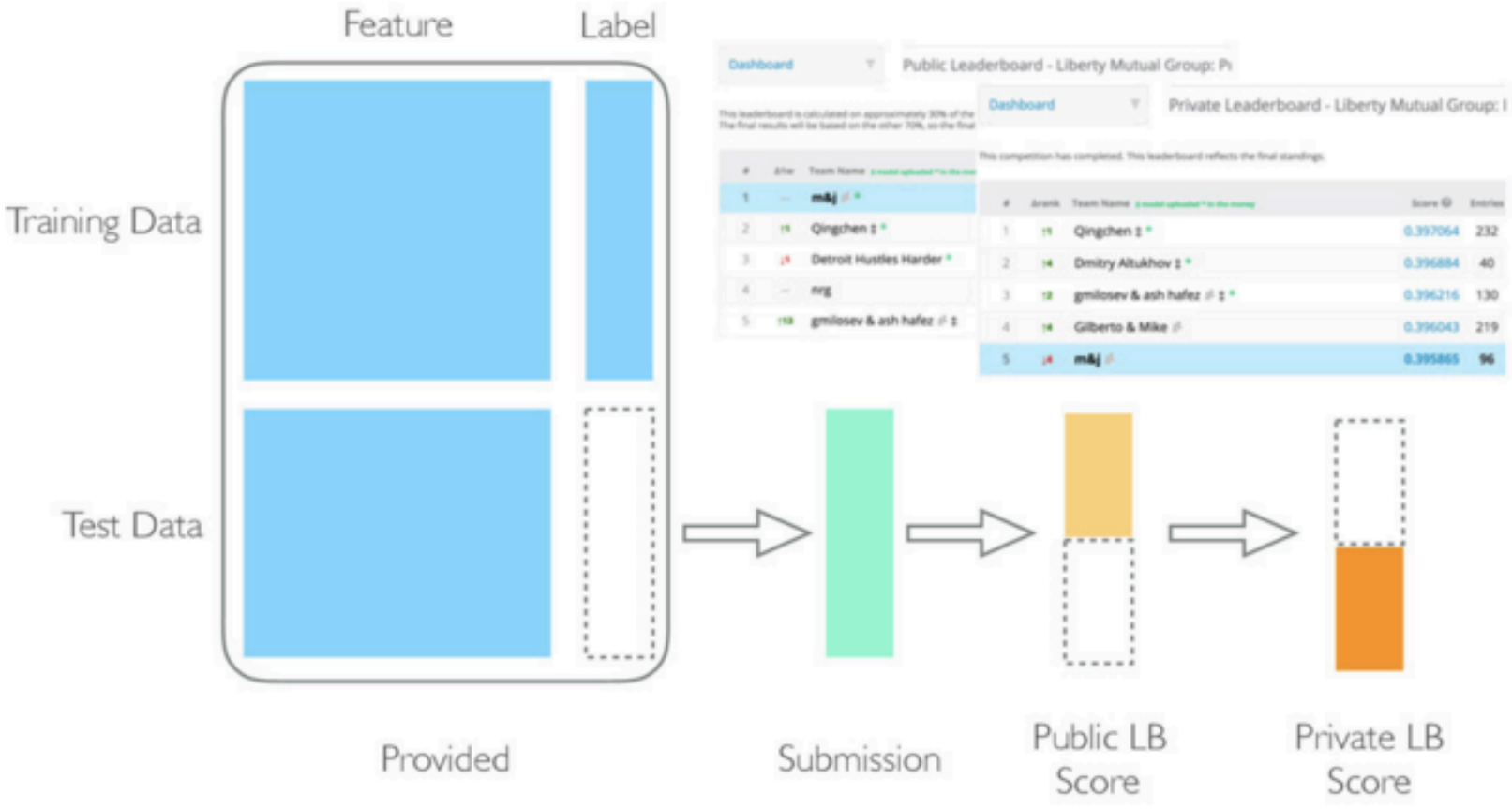
# Kaggle Competition

- Why use Kaggle?
  - Hard to find resources
  - Kaggle
    - Excellent platform
    - Runned by skilled professionals
    - Community is very active, collaborative and knowledgeable

# Common Competition Challenges

- Data preparation
  - Create new variables, reformatting variable values
- Missing values
  - Missing value indicators, record deletion
- High-level categorical variables
  - Reduce number of levels
- Combining multiple datasets
  - Decide which fields are useful, join records by certain columns
- Dimensionality reduction (feature selection)
- Need for ensembles
  - One model may not be enough
- Importance of decimal places
  - Every decimal place of improvement is crucial in these competitions where modelers are submitting hundreds of entries

# Competition Structure



# Best Practices

# Best Practices

- Feature Engineering
- Machine Learning
- Cross Validation
- Ensemble

# Feature Engineering

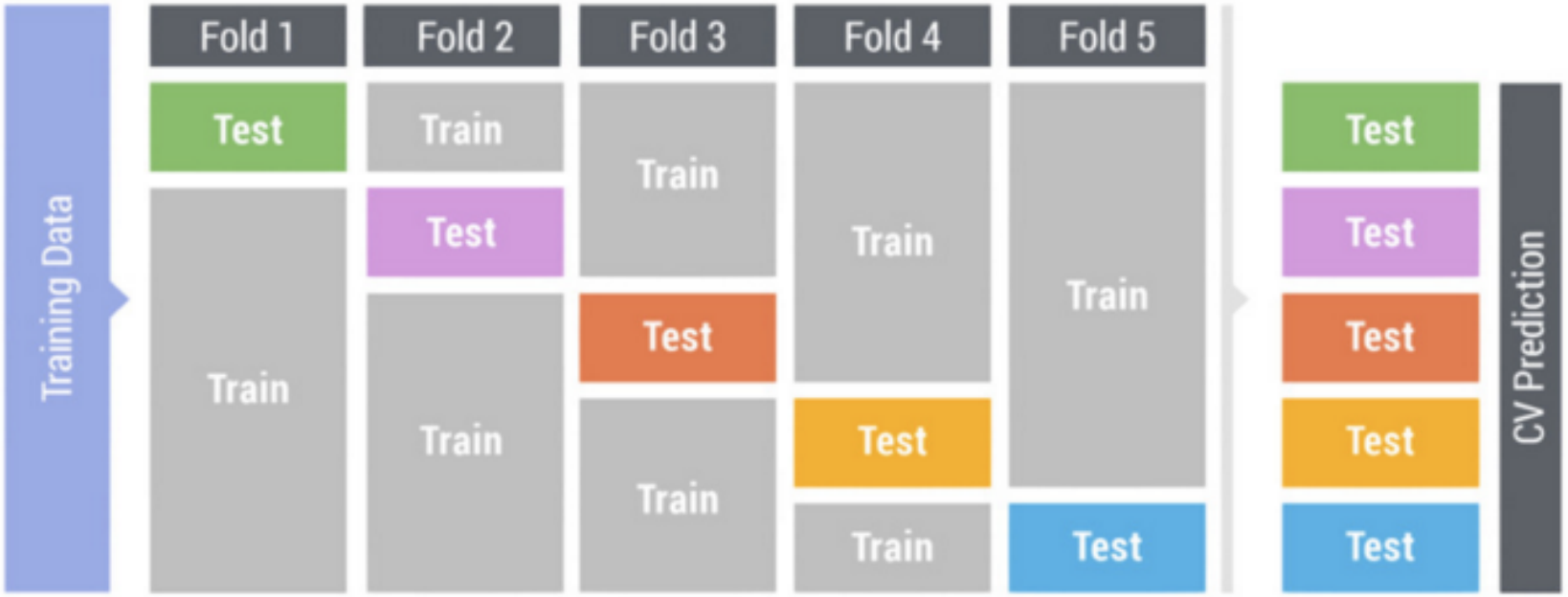
- Numerical - Log, Log(1+x), Normalization, Binarization.
- Categorical – One-hot-encode.
- Timeseries – Stats, FFT
- Numerical/Timeseries to Categorical –RF/ GBM*

http://www.csie.ntu.edu.tw/~r01922136/kaggle-2014-criteo.pdf

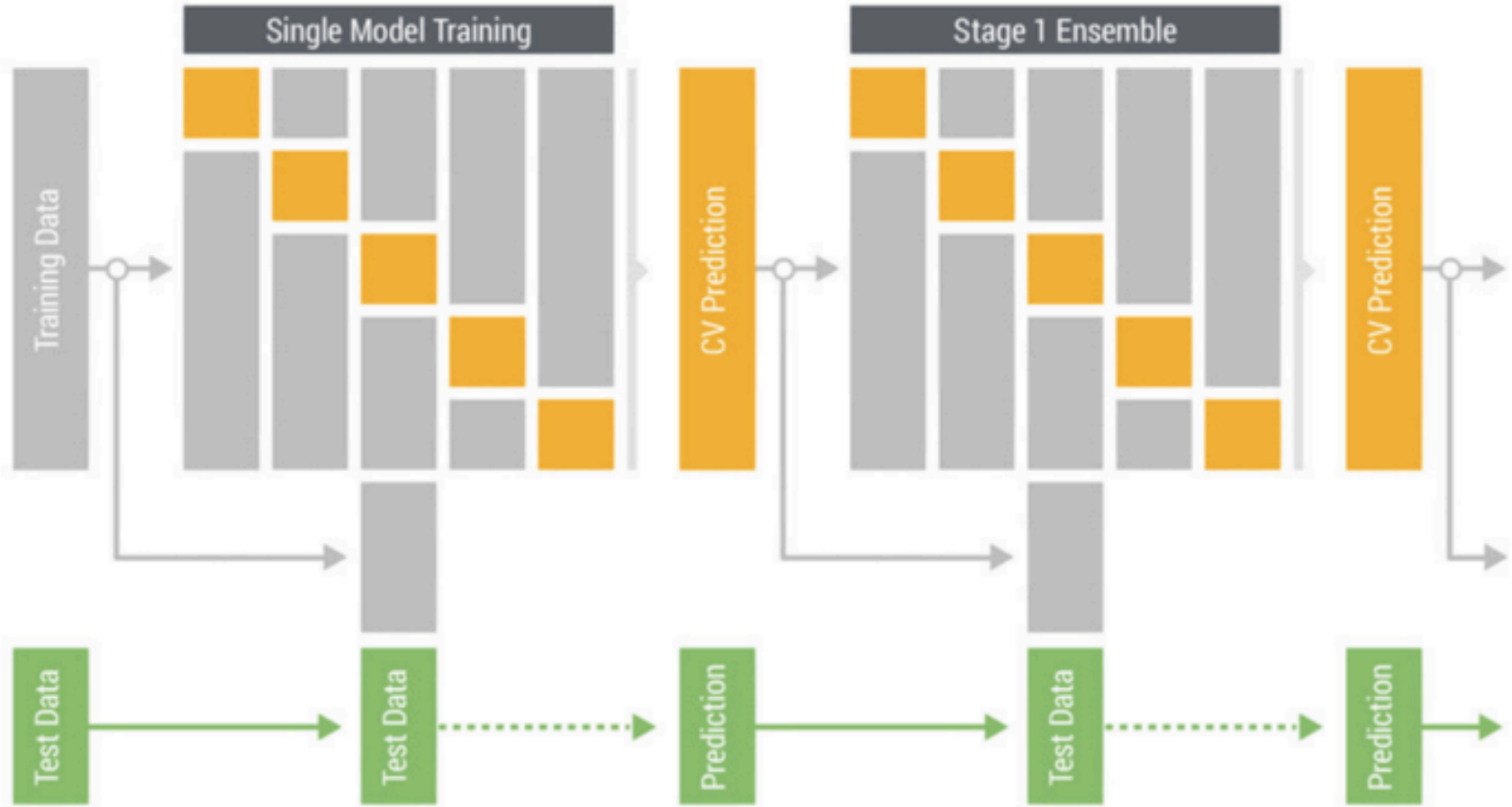
# Machine Learning

| Algorithm | Tool | Note |
|---|---|---|
| Gradient Boosting Machine | XGBoost | Best out-of-the-box |
| Random Forests | Scikit-learn, randomforest | |
| Extra Trees | Scikit-Learn | |
| Regularized Greedy Forest | | |
| Neural Networks | Caffe, Keras | Blend with GBM and for image recognition. |
| Logistic/Linear Regression | Scikit-Learn, Vowpal Wabbit | Fastest, Good for ensemble. |
| FTRL | Vowpal Wabbit | CTR estimation |
| Factorization Machine | libFM | KDD cup 2012 |
| Field-aware Factorization Machine | libFFM | CTR estimation competitions |
| Support Vector Machine | SVM, scikit-learn | |

# Cross Validation

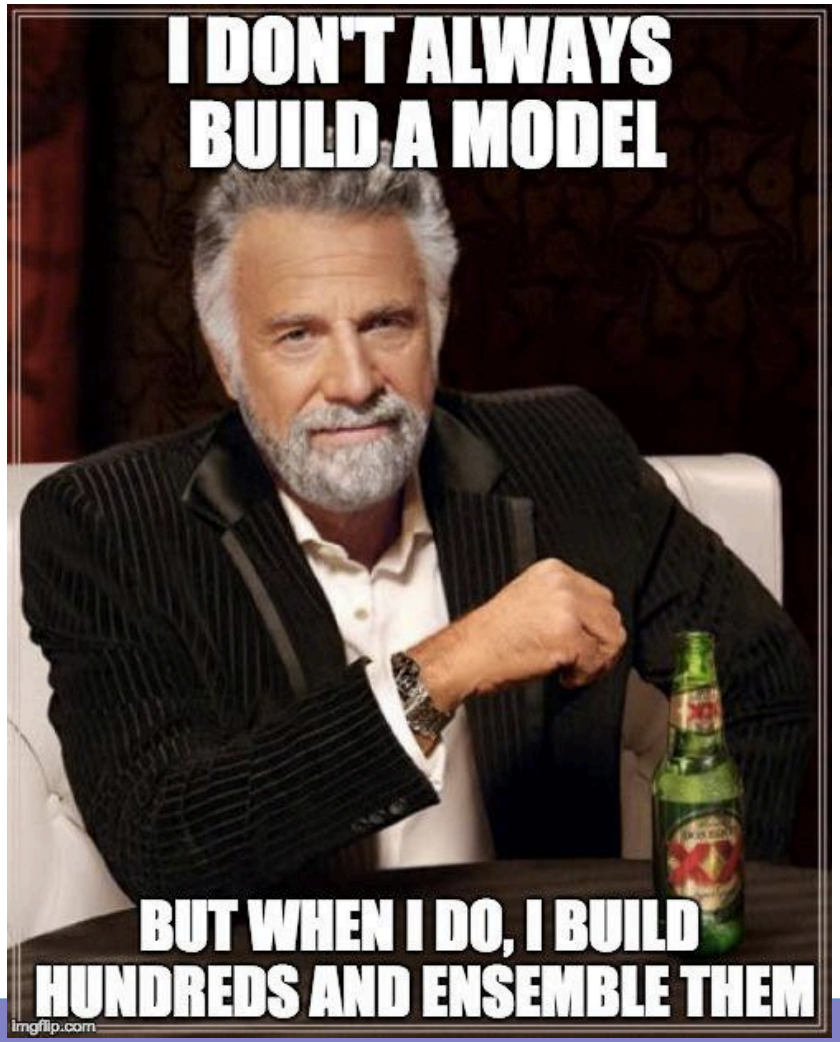


# Ensemble




http://mlwave.com/kaggle-ensembling-guide/

# Ensemble

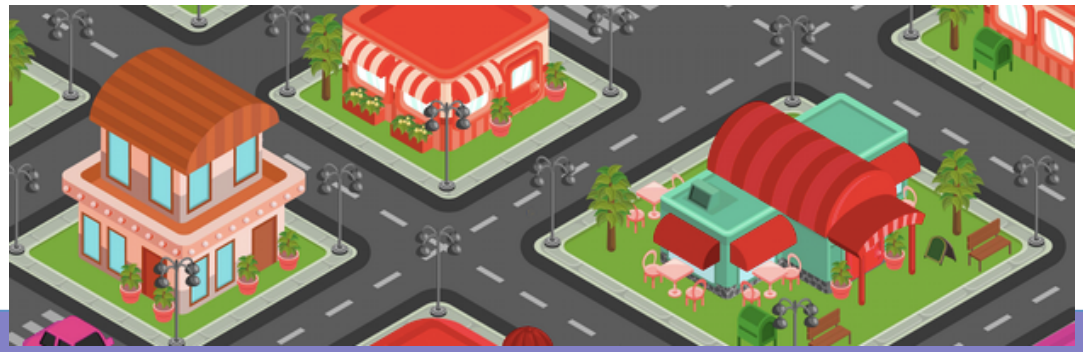

I DON'T ALWAYS
BUILD A MODEL
BUT WHEN I DO, I BUILD
HUNDREDS AND ENSEMBLE THEM
imgflip.com

# Example: KDD CUP 2015

# Restaurant Prediction: Problem

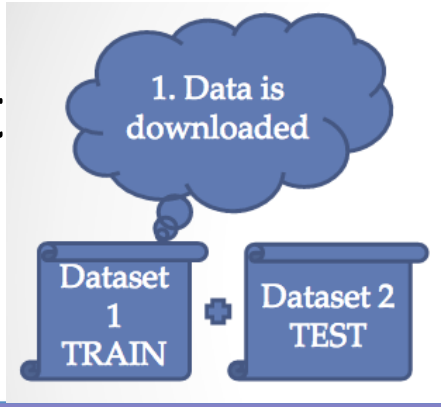
- TFI is the company behind: Burger King, Sbarro, Popeyes, Usta Donerci, and Arby's.
- Employ around 20,000 people around Europe and Asia
- Where and When to open new Restaurants.
- If wrong location is chosen, the restaurants closes within 18 months and operating losses are incurred
- Find a mathematical model to increase the effectiveness of investment in new restaurants.

TFI
TAB
FOOD
INVESTMENTS



# Data

- Training set consists of dataset with 137 restaurants.
- Test set consists of 100000 restaurants.
- Data consist of columns open date, location, city type, demographic data, real estate data and commercial data.
- Target column is revenue of the rest



# Data Fields

- Id : Restaurant id.
- Open Date : opening date for a restaurant
- City : City that the restaurant is in. Note that there are unicode in the names.
- City Group: Type of the city. Big cities, or Other.
- Type: Type of the restaurant. FC: Food Court, IL: Inline, DT: Drive Thru, MB: Mobile
- P1, P2 - P37: There are three categories of these obfuscated data. Demographic data are gathered from third party providers with GIS systems. These include population in any given area, age and gender distribution, development scales. Real estate data mainly relate to the m2 of the location, front facade of the location, car park availability. Commercial data mainly include the existence of points of interest including schools, banks, other QSR operators.
- Revenue: The revenue column indicates a (transformed) revenue of the restaurant in a given year and is the target of predictive analysis.

# Evaluation

- Root mean squared error of the test sets.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

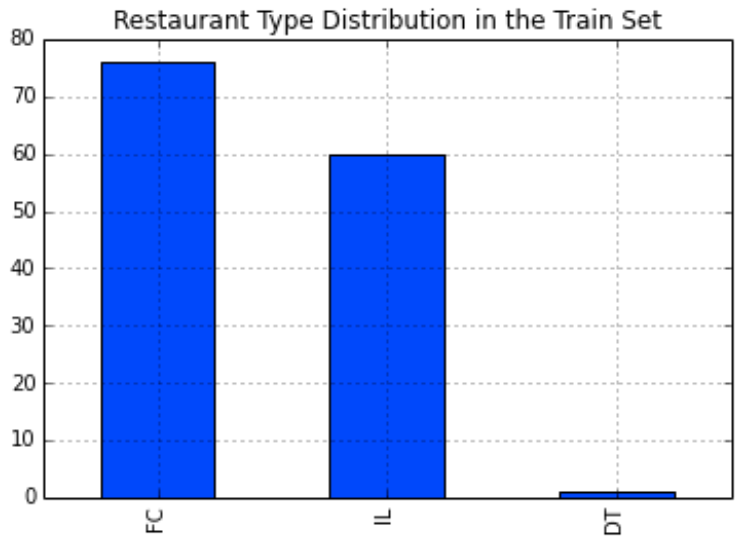
Where y-hat$_i$ is the predicted revenue of the i$^{th}$ restaurant and y is the actual revenue of the i$^{th}$ restauarant.

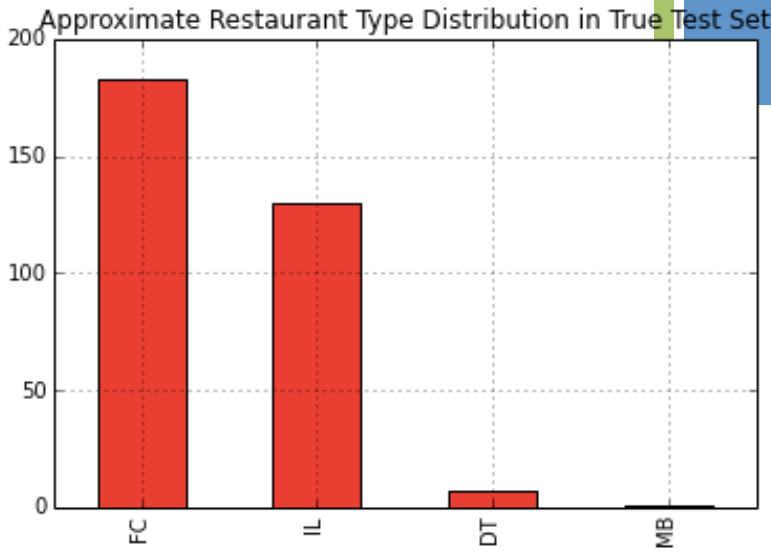
# Code R

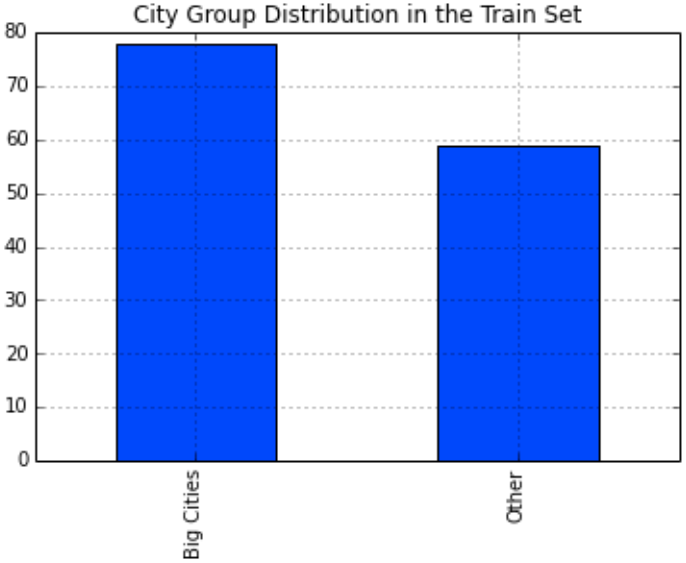
```
library(party)
#path of folder
Setpwd("/Users/bikash/repos/RestaurantRevenuePrediction/")
#load data
train = read.csv(”data/train.csv", header = TRUE, stringsAsFactors = FALSE)
test = read.csv(”data/test.csv", header = TRUE, stringsAsFactors = FALSE)
head(train)
View(train)
```

| P26 | P27 | P28 | P29 | P30 | P31 | P32 | P33 | P34 | P35 | P36 | P37 | revenue | days_open | weeks_open | years_open |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 4.0 | 2.0 | 3.0 | 5 | 3 | 4 | 5 | 5 | 4 | 3 | 4 | 5653753 | 5728.0417 | 818.29167 | 15.693265 |
| 0.0 | 0.0 | 3.0 | 3.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6923131 | 2594.0000 | 370.57143 | 7.106849 |
| 0.0 | 0.0 | 1.0 | 3.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2055379 | 744.0000 | 106.28571 | 2.038356 |
| 2.5 | 2.5 | 2.5 | 7.5 | 25 | 12 | 10 | 6 | 18 | 12 | 12 | 6 | 2675511 | 1145.0000 | 163.57143 | 3.136986 |
| 3.0 | 5.0 | 1.0 | 3.0 | 5 | 1 | 3 | 2 | 3 | 4 | 3 | 3 | 4316715 | 2144.0417 | 306.29167 | 5.874087 |
| 0.0 | 0.0 | 7.5 | 5.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5017319 | 1865.0000 | 266.42857 | 5.109589 |
| 4.0 | 5.0 | 1.0 | 3.0 | 4 | 5 | 2 | 2 | 3 | 5 | 4 | 4 | 5166635 | 1624.0417 | 232.00595 | 4.449429 |
| 0.0 | 0.0 | 3.0 | 2.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4491607 | 1371.0417 | 195.86310 | 3.756279 |
| 4.0 | 2.0 | 2.0 | 3.0 | 4 | 5 | 5 | 3 | 4 | 5 | 4 | 5 | 4952497 | 1668.0417 | 238.29167 | 4.569977 |
| 0.0 | 0.0 | 5.0 | 2.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5444227 | 1223.0000 | 174.71429 | 3.350685 |
| 0.0 | 0.0 | 10.0 | 2.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3745135 | 591.0417 | 84.43452 | 1.619292 |

# Feature Engineering



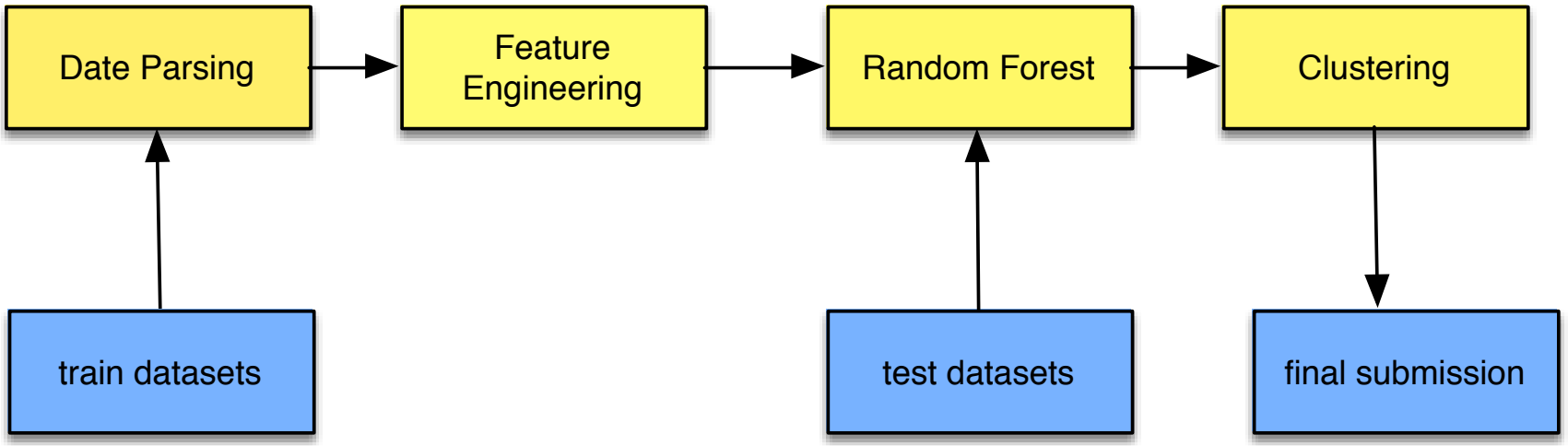



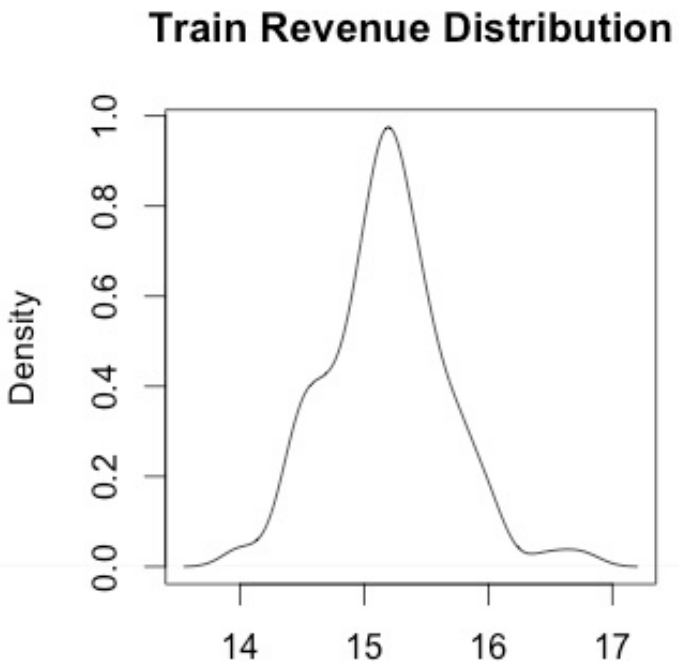
# Model Setup

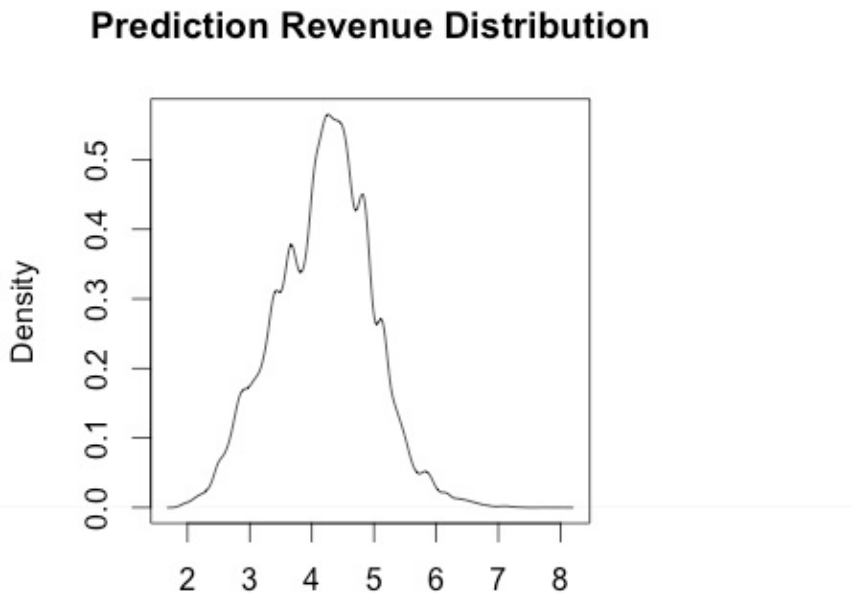
- Prior to model selection, we add a constraint to restrict Revenue outliers greater than 10million.
- We do no transformations of any variables other than to create an integer variable of Years Open from Date. We set the variable Revenue as the Target.
- From the list of potential candidate variables, we aim to reduce this by about half(seeking a more parsimonious model) but we can simply exclude some variables that either have too many levels (i.e., Cities) or too few levels (i.e., City Group).
- The only default selection that we change is to change our testing method to 'No independent testing –exploratory model', given that we have an insufficient quantity of cases to make any other type of testing method valid in this particular context.

# Model


Date Parsing
Feature Engineering
Random Forest
Clustering
train datasets
test datasets
final submission

# Model


Train Revenue Distribution
Density
0.0
0.2
0.4
0.6
0.8
1.0
14
15
16
17


Prediction Revenue Distribution
Density
0.0
0.1
0.2
0.3
0.4
0.5
2
3
4
5
6
7
8

# Private Leaderboard

| # | Δrank | Team Name ‡ model uploaded * in the money | Score | Entries | Last Submission UTC (Best – Last Submission) |
|---|---|---|---|---|---|
| 1 | ↑205 | Arsenal ‡ * | 1727811.48554 | 21 | Fri, 24 Apr 2015 03:14:27 (-23.9d) |
| 2 | ↑42 | Climent_Josep ‡ * | 1729265.36129 | 27 | Mon, 04 May 2015 06:32:55 (-28.2h) |
| 3 | ↑338 | OldSchool ‡ * | 1732292.38960 | 72 | Tue, 28 Apr 2015 08:53:55 (-3.9d) |
| 4 | ↑170 | McData | 1734366.17757 | 53 | Fri, 01 May 2015 06:01:48 |
| 5 | ↑294 | Statsfreak | 1745878.25649 | 13 | Fri, 01 May 2015 00:15:30 |
| 6 | ↑61 | Manuel Díaz (TFI) | 1748986.04032 | 74 | Mon, 04 May 2015 08:37:36 (-25.2h) |
| 18 | ↑174 | AndrewH | 1765397.73448 | 36 | Fri, 01 May 2015 19:01:25 (-4.8d) |
| 19 | ↑612 | ash hafez | 1765819.71675 | 29 | Mon, 04 May 2015 01:31:15 (-5.5d) |
| 20 | ↑13 | yonidahan | 1766973.99205 | 80 | Mon, 04 May 2015 16:00:42 (-2.2h) |
| 21 | ↑81 | Gabin | 1767532.94727 | 54 | Thu, 30 Apr 2015 11:29:44 (-0.1h) |
| 22 | ↑158 | wahahawoohoo | 1767818.49896 | 6 | Mon, 27 Apr 2015 19:30:52 (-4d) |
| - | | **Bikash Agrawal** | **1768307.83986** | - | **Tue, 03 May 2016 12:53:11** Post-Deadline |

**Post-Deadline Entry**
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

| | | | | | |
|---|---|---|---|---|---|
| 23 | ↑277 | Ujjwal Karn | 1768663.01295 | 36 | Mon, 04 May 2015 17:17:02 (-3.8d) |

# Conclusion

- There will usually be some degree of data preparation needed
- Inconsistencies between training and testing samples should be expected
- The best model will never be your first - automation is your friend in building many different models with adjusted parameters
- You can learn a lot from resubmitting models to the leaderboard and adjusting accordingly

# How can kaggle help you?

- Kaggle help you to get job if you do good in the competition.
- Good award if you win the competition.

# Thank you

bikash@boost.ai

# Case 2: Titanic

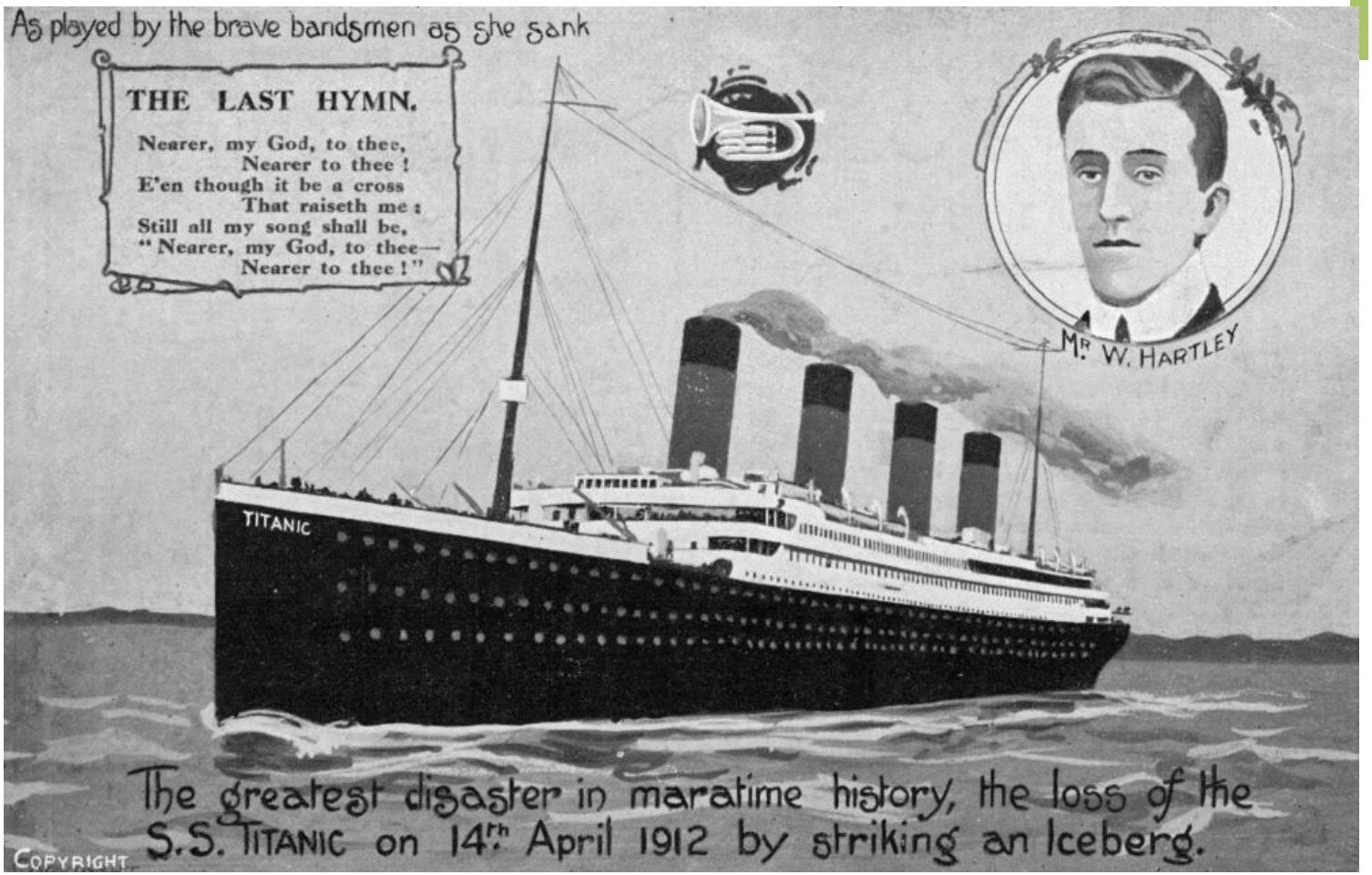
- https://www.kaggle.com/c/titanic

**Titanic: Machine Learning from**

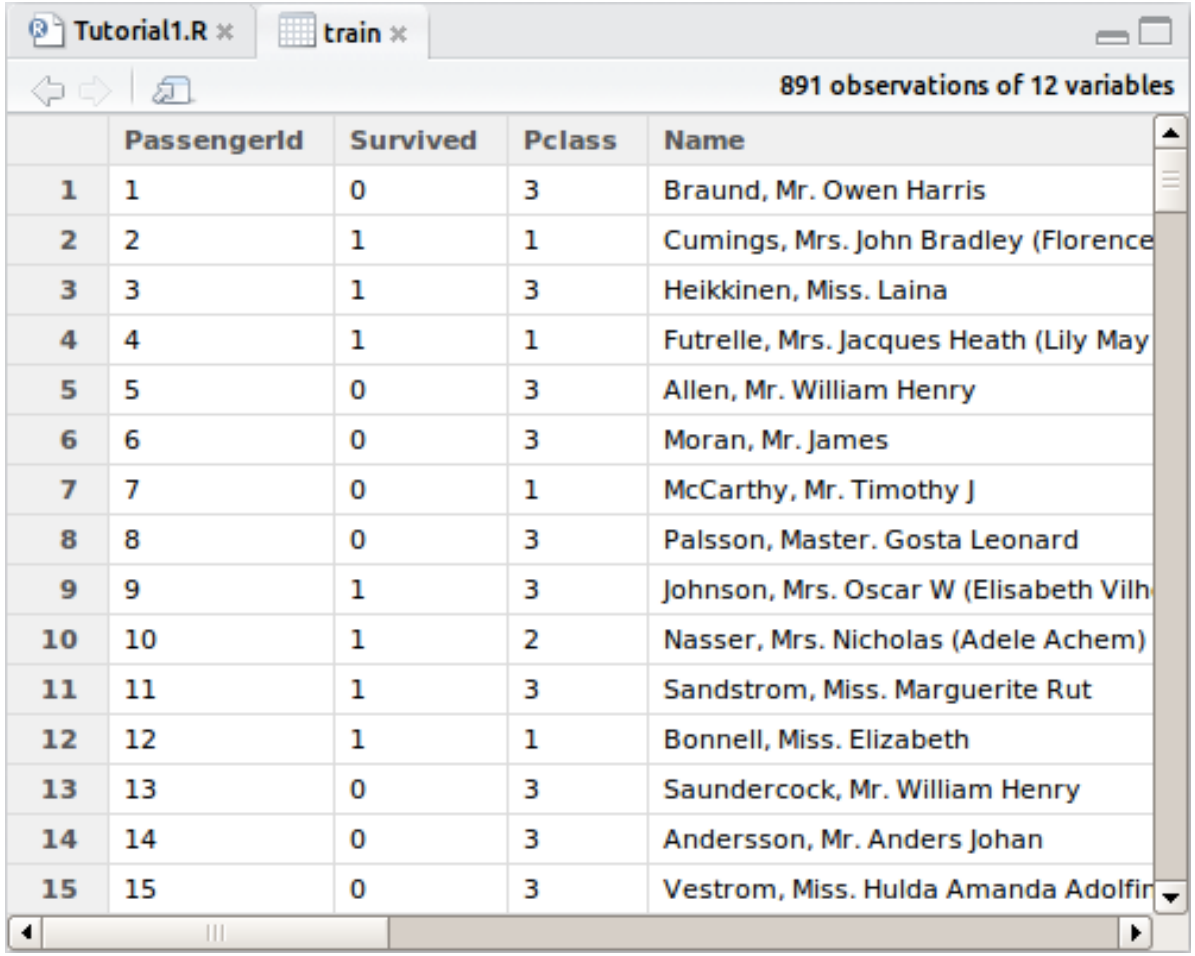
Predict survival on the Titanic using Excel, Python, R & Random Forests

# Introduction

- The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.


As played by the brave bandsmen as she sank
THE LAST HYMN.
Nearer, my God, to thee,
Nearer to thee!
E'en though it be a cross
That raiseth me;
Still all my song shall be,
"Nearer, my God, to thee—
Nearer to thee!"
Mr W. Hartley
TITANIC
The greatest disaster in maratime history, the loss of the S.S. TITANIC on 14th April 1912 by striking an Iceberg.
COPYRIGHT

# Data


Tutorial1.R × train ×

891 observations of 12 variables

| | PassengerId | Survived | Pclass | Name |
|---|---|---|---|---|
| 1 | 1 | 0 | 3 | Braund, Mr. Owen Harris |
| 2 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence |
| 3 | 3 | 1 | 3 | Heikkinen, Miss. Laina |
| 4 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May |
| 5 | 5 | 0 | 3 | Allen, Mr. William Henry |
| 6 | 6 | 0 | 3 | Moran, Mr. James |
| 7 | 7 | 0 | 1 | McCarthy, Mr. Timothy J |
| 8 | 8 | 0 | 3 | Palsson, Master. Gosta Leonard |
| 9 | 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilh |
| 10 | 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) |
| 11 | 11 | 1 | 3 | Sandstrom, Miss. Marguerite Rut |
| 12 | 12 | 1 | 1 | Bonnell, Miss. Elizabeth |
| 13 | 13 | 0 | 3 | Saundercock, Mr. William Henry |
| 14 | 14 | 0 | 3 | Andersson, Mr. Anders Johan |
| 15 | 15 | 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfin |