Kaggle: A Platform For Making Data Science As a Sport

Institute of Computer Science
Georg-August-Universität Göttingen

Advisor: Narisu Tao
Lecturer: Prof. Dr. Xiaoming Fu
Author: Hari Raghavendar Rao Bandari
11334055
h.bandari@stud.uni-goettingen.de
Date: April 1, 2015

Abstract

The web is full of data-driven application, In every field of business generates data and analyzing that data is complicated. If we solve this particular data problem and find new features and patterns in the existing data, then we can create a new "data product" for that it requires knowledge about the data. Due to increase in the volume and details of information captured by enterprises, and rise, of multimedia, the internet things will fuel exponential hike in the data. In fact the growth rate of the data is increasing from megabytes to Zetabytes with this data scientist workload will be increased and it will be complicated to understand the unstructured data. Now to solve this unstructured messy data, we need a data scientist, so Kaggle is helping companies and researchers post their data and problem, then all the data, scientists compete against each other competitors in teams or individually to solve a particular complicated data science problem. To experience data science problem I participated in "digit recognizer" competition in kaggle and shared my experience at the end of the report.

Contents

1 Introduction

The web is full of data-driven application. Almost in every field of business generates data and analyzing that data can improve the performance of the business, but analyzing data is a complicated task, we have to apply a different variety of techniques to get value added product to the business. Today's technology is helping in every field of business with this volume of the data is increasing and storage capacity also increasing and which leads to unstructured data. So business has a database behind a web front end, and middleware that communicates to a number of other databases and data services. To analyze the complicated data and finding a value added product from the existing data will be a new profit market for the business. We need dedicated people who can analyze and manipulate data to derive insights and build data products. It combines skill-sets ranging from computer science, to mathematics, to art.

2 What is Data Science

Data science is an extraction of knowledge from data. It applies different techniques and theories drawn from many fields within the board areas of mathematics, statistics, information theory and information technology including machine learning, pattern recognition, predictive analytics.

Data science enables the creation of "data products". For example "Google" created a data product from their data itself. Google "PageRank" algorithm was among the first to use data outside of the page itself, in particular, the number of links pointing to a page. Tracking links made Goggle search engine more efficient and fast.

2.1 Where data comes from

Data is generated by government, webserver, small business firms, even our body. Now a days people spending lots of leisure time online, and leaving a trail of data wherever they go such as Mobile application leave an even richer data trail, Since many of them uses geolocation feature to know user location details which is very much useful for predicting actions and recommending them nearest hotels and

shopping malls , All this data would be useless if we couldn't store it.

2.2 Why we are talking about the data Science currently

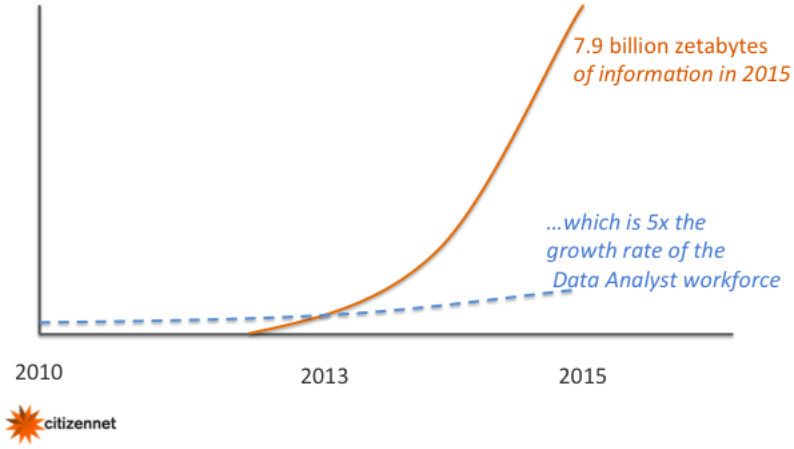

Figure 1: shows the growth rate of the data Analyst Workforce[1]

figure 1 Explains, The amount of the data in our world has been exploding and analyzing large data sets so called as big data. The increasing volume and details of information captured by enterprises, the rise of multimedia and the internet of things, will fuel exponential hike in the data and above graph, explain about how the data size is increased from megabytes to zeta bytes. In fact the growth rate of the data is conservatively 5x the projected growth rate of the data analyst workforce.

3 Data Scientist

3.1 The sexiest job of the 21st Century

Harvard Business Review puts the data scientist into the national limelight in their publication, as the "Data Scientist: The sexiest job of the 21st Century". As the popularity for the data scientist position got attention and at the same time Mckinsey estimates 140,000 to 190,000 shortage by 2018.

[1]https://citizennet.com/blog/2014/01/03/four-trends-in-marketing-technology-for-2014/

"The shortage of data Scientist is becoming a serious constraint in some sectors"

For every Big data problem, the solution often rests on the shoulders of the data scientist. The role of the Jonathan Goldman as a data scientist has changed the status of the LinkedIn company with his analytical skills and Reid Hoffman, LinkedIn co-founder and CEO at that time had faith in the power of analytics because of his prior experiences of pay pal company and he had granted Jonathan Goldman a high degree of autonomy.

3.2 How Data Science help to stretch boundary of human knowledge

"Machine Intelligence Cracks Genetic Control" The researchers from a "computational biologist" at the University of Toronto revealed the differences between the Autism person and Healthy human because genetic researchers do research on the 1 percent of the genome the areas that code for protein but in reality whole genome is important for autism research. Data Scientists can resolve this kind of problem if they have access to a whole set of data problem.

3.3 What kind of skills are owned by Data Scientist

According to "2011 EMC Survey" proudly points out that 40 percent of data scientists had masters degrees or better. Many data specialists such as "election forecaster Nate Silver", "Moneyballs Paul DePodesta" and "Clouderas Jeff Hammerbacher" among them only have "bachelors degrees". Data Scientist skill sets into two categories. They are 1. Technical Skills 2. Business Knowledge

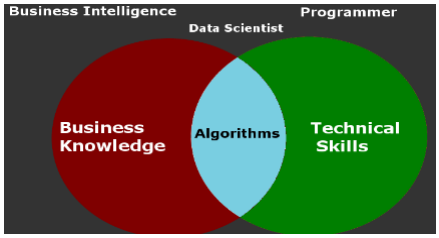

Figure 2: Pie diagram of data scientist skills[2]

Data Scientist should have technical skills who can program and should have business Intelligence who have internal business knowledge, but not mandatory with these skills data Scientist should able to write an algorithm such that complicated problems should resolve. Technical Skills again categorized into two Types. They are

3.4 Technical Skills

Theoretical Knowledge

Mathematics

Basic undergraduate courses cover calculus and linear algebra. Once basics are strong, its time to go deep into matrix computation, diffusion geometry, and similar topics in applied mathematics. Most data mining applications use matrix computations as their fundamental algorithms, so a strong understanding of them is essential.

Statistics

Understanding correlation, multivariate regression and all aspects of massaging data together to look at it from different angles for use in predictive and prescriptive modeling is the backbone knowledge thats really step one of revealing intelligence. Statistical tools: R, SAS, SPSS, Scipy, Stats.

Data Mining

Data mining is about big data sets to discover new and interesting patterns. E.g., Cluster analysis, anomaly detection, dependencies, supervised algorithms-Naive Bayes, Decision Tree, Neural Network, etc. And non-supervised -Association Rules, Clustering, etc.

Data Modelling

The process used to define and analyze data requirements needed to support the business processes within the scope of corresponding information systems in or-

ganizations. E.g., Agile Data Modeling, ORM Modeling, UML class diagrams, ERWin, CRC cards etc.

Predictive Modelling

It is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. To test skills, take part in Kaggle competition.

Machine Learning

"Ability of a machine to improve its own performance through artificial intelligence" and "Use of computers to develop and improve algorithms".

Visualization

Showcasing the finding features or patterns in a way that business users understand it. E.g., Flare, HighCharts, amCharts, Google visualization API etc.

Practical Skills

Programming /Scripting Languages

A scripting language is a High level programming language written for a special run-time environment that can interpret and automate the execution of tasks that could alternatively be executed one-by-one by a human operator. E.g., python, C/C++, Java, Ruby, Perl, MATLAB, Pig etc.

Distributed Computing Systems

It is a computing concept that refers to multiple computer systems working on a single problem. In distributed computing, a single problem is divided into many parts, and each part is solved by different computers. As long as the computers are networked, they can communicate with each other to solve the problem. E.g., Hadoop, Hbase, Cassandra, MapReduce, Hive etc.

Relational Databases

A solid understanding of SQL-based systems is required. Should have the fundamentals of database design and management. Get a handle on primary and foreign keys, indexing, querying, normalizing, constraints. NewSQL: Cloudera Impala, Clustrix, VoltDB, etc.

3.5 Business Knowledge

Domain Expertise

This means having deep knowledge and interest in your field of expertise and total understanding of your company data. E.g., medicine, government, retail, manufacturing, etc.

Creativity and Curiosity

Creative data scientists are optimists. Every day, they look at a hodgepodge of incomplete and messy data, inadequate analytic, faulty methods and models, and seemingly insoluble business problems. Creative data scientists are curious. They arent afraid of playing around in unstructured environments, of proceeding by trial and error, of following the white rabbit down the hole.

Storytelling

Once the data scientist finds new patterns or feature in the messy data they have to narrate a story so that others in the company can understand it. So it's like selling your thoughts and innovation to the business people.

Project Management

It's about leadership position, pressure and have to grasp the wider context, delegate tasks, is conscious of budget and time constraints and play to the strengths of the folks working for you.

Ethics

These are risks your company cant afford to ignore. Whatever the project, youll need to consider what effect your work will have on the lives of customers and clients.

The Elevator Speech

An advanced degree in a quantitative field hands-on experience hacking data, good exploratory analysis skills, the ability to work with engineering teams, and the ability to generate and create algorithms and models rather than relying on out-of-the-box ones.

4 Kaggle Platform

4.1 Netflix Competition

Netflix hosted an open competition in 2009 for best collaborative filtering algorithm for predicting user rating for films and competition was held on online DVD rental and online video streaming service. On 21 September 2009, the grand prize of US $1,000,000 was given to the "BellKor's Pragmatic Chaos team" which bested Netflix's own algorithm for predicting ratings by 10.06%. To improve Netflix predicting rating by 10.06% so many data scientists were participated and revealed many solutions for one problem in less period of time with different approaches to improve the accuracy of the user predicting rating. So here Netflix itself hosting their problem with the data. But with the rise of data scientist, Kaggle has been motivated to start hosting a competition of different companies' problems under one platform where all data, scientists around the world can participate in the competition and win competition prizes.

4.2 Introduction

kaggle is a website, which was founded by Anthony Goldbloom in 2010 in Melbourne, and moved to San Francisco in 2011. In November 2011, Kaggle announced Series A funding led by Index Ventures and Khosla Ventures and now it

is the worlds largest community of data scientists. All the data scientists compete with each other in teams or individually to solve a particular complicated data science problem, and the top competitors are invited to work on most challenging, delicate business complicated issues with some of the top leading companies through Masters Competition. kaggle provides leading edge data science results to companies of all sizes. It has an excellent track record of solving real-world complex problems across a distinct bunch of industries including life sciences, financial services, energy, information technology and retail. A kaggle community of data scientists includes 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 IT technology. The competitors from over 100 countries and 200 universities. In addition to the prize money and data, they use Kaggle to meet, learn, network and collaborate with experts from related fields. Participation of Kaggle competitions across the world.

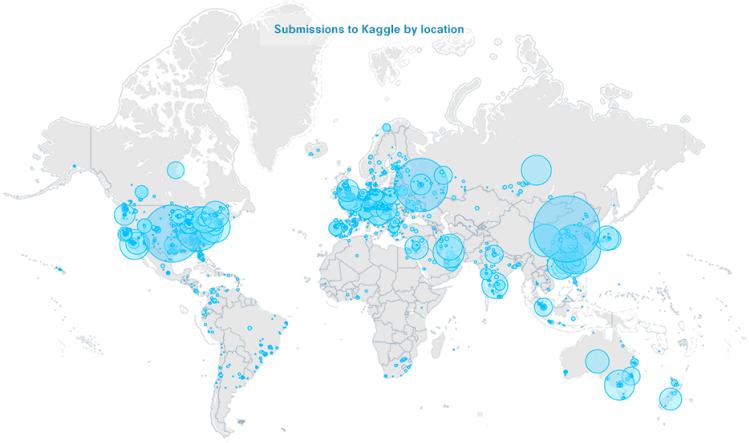

Figure 3: Submission to kaggle by location[3]

4.3 Kaggle Business Revenue

Kaggle is a having two margin market value that links the gap between data problems and data solutions, It's doesnt charge any fees to participants who participate in the competition, complete fees are paid by the owner of the data problem. It

[3]https://www.kaggle.com/about

business began as a platform for hosting public data science challenges, in which sponsors post their complex problem to the Kaggle platform, then data, scientists from all over the world compete in the respective challenge to create the best solution for the data problem. It collects a hosting fee from the competition sponsor and may work as a consultant to help challenge the structure and prepare the dataset. Data scientists who have solved the particular data science problem in the Kaggle competition will get hired by the top most leading companies as the "Data Scientist" in their respective company.

4.4 How Kaggle Competition Works

1.Upload a prediction Problem

Figure 4: Upload a Prediction Problem[4]

The competition host provides the raw data with the description of that problem to Kaggle. Kaggle offers a consultancy service to prepare the data, from the competition, anonymize the data, and set the rules for the challenge. And incorporate the best model into their operations.

2.Submit

Depending upon on the challenge duration of the competition, several participants experiment with different statistical techniques and compete against each other in the form of teams or individuals depending upon on the competition prize money and the complication of the problem and to produce the best models with high predictive accuracy. Their submissions are scored immediately (based objectively on their predictive accuracy, relative to a hidden solution file) and summarized on a live Leaderboard for who is leading the competition.

[4]http://plugg.com.au/main/article/kaggle-where-phd-level-people-help-and-compete-each-other-solve-problems

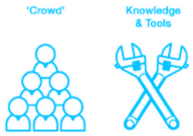

Figure 5: submit solution[5]

3.Evaluate and Exchange

Figure 6: Evaluate and Exchange[6]

Once the deadline finishes, the competition host organization pays prize money in exchange for the winning model or models.

4.5 Kaggle Top Competition

1: Uploaded a prediction problem by NASA and Royal Astronomical Society

The challenge was about a cosmological image analysis to measure the small distortion in galaxy images caused by dark matter. The Competition prize money is an expense paid visit to the NASA Jet Propulsion Laboratory (JPL). Industry Domain: Astronomy, Data Type: 100,000 PNG images simulated galaxies with burring Task: Find ways to measure the true galaxy shape (e1 and e2 values).

2: Submit

In this competition almost 75 players on 70 teams and submitted 760 entries length of the competition was 3 month within three weeks Martin O'Leary, a

[7]https://www.kaggle.com/c/mdm

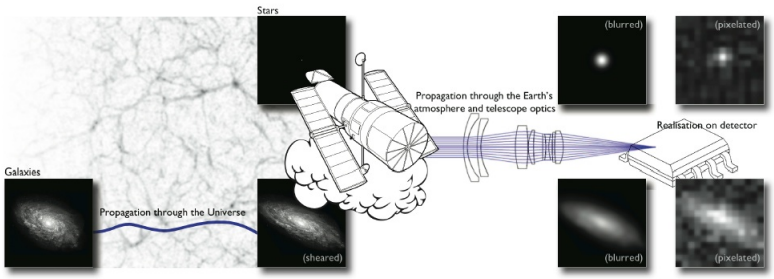

Figure 7: Galaxy Images[7]

British glaciologist, had created a solution using "outperformed the state-of-the-art algorithms". Meanwhile, David Kirkby and Daniel Magala, cosmologists at UC Irvine, developed an "Artificial neural network" to recognize patterns in the galaxy images their code was written in C++.

3: Evaluate and Exchange

Competition deadline, finished and the winning team solution was 3x increase in accuracy in state of art benchmark that had taken NASA decades to develop. The winning team "Deepzot"(David Kirkby and Daniel Magala) were awarded a trip to present their methods to NASA and other agencies.

4.6 Kaggle: User ranking and Tier System

Kaggle has two categories system for measuring community member's performance based on Ranking active users based on points earned through competition and a tier system based on performance of the users. Points: Kaggle users awarded with points according to their performance in the competitions. These points are recalculated at the end of each competition once results has been finalized because until the deadline finishes the users can submit his best predictive model accuracy or Kaggle it selects by default his respective users' best accuracy model solution based on his submit. If users are participating in teams their points are calculated on the current formula for each competition splits the points among the team members, decays the points for lower finishes, adjusts for the number of teams that entered the competition, and linearly decays the points to 0 over a two-year period from the end of the competition. For each competition, the formula is:

```
(100000/# Team Members) (Team Rank)^0.75
(log10(# Teams)(2 years-time since deadline/2 years))
```

Points of the each users are calculated with the time decay fixed at the time of the most recent competition deadline. Tier: Kaggle users are divided into three sectors they are

NOVICE

A New user just registers with Kaggle website and not yet participated in any competition and with no ranking points.

KAGGLER

This stage is a wildest tier, illustrating the main body of the Kaggle community. This stage of user can have low ranking points. Members in this group should have 5000 points from competition performance to have access to the private messaging system to contact other Kaggler users on the site. So they can form teams or seeking suggestions from higher ranking kaggler users are trying to know the solutions of the competition in which respective kaggler ranked topped and this stage is of mixed ranking users.

MASTER

To achieve this tier user must fulfil two criteria. They are Consistency: at least 2 Top 10% finishes in public competitions. Excellence: at least 1 of those finishes in the top 10 positions. They are few categories competitions which cannot be counted towards earning Master tier status and they are general competitions which are held for teaching and practicing include Kaggle-in-class competitions, and Hackathons.

Highest Rank Ever

The current user rank will be calculated by iterating through all historical competitions deadlines, calculating the rankings at those times and then finding the best rank for each user. The highest rank ever achieved is also shown on the user profile. In total, participants invest hundreds of hours in exploring the exponential

solution to the given problem. Teams are constantly challenged each other in the forms of teams or individuals depending upon the competition and leapfrog until the high score until the competition deadline. This below visualization shows participants ranking over the course of the competition, here below each line represent one team.

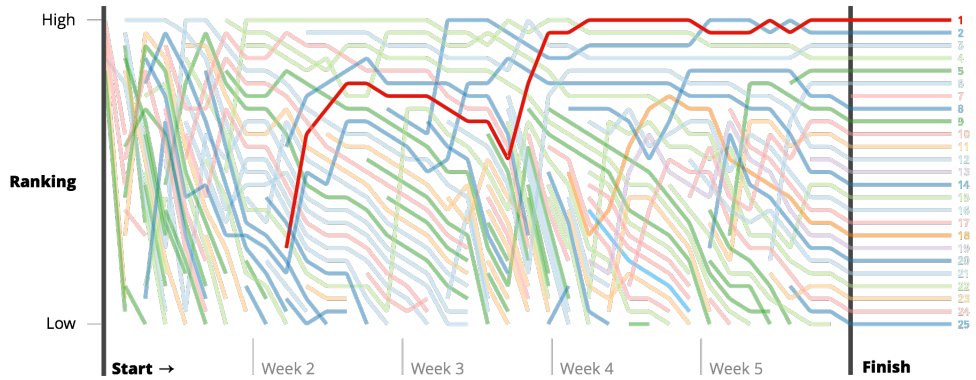

Figure 8: Weekly team Ranking[8]

4.7 Kaggle: private leaderboard vs public Leaderboard

Public Leaderboard

Kaggle competitions are judged by your model performance on a test data set. Kaggle has the solution for the particular data set, but they compare with their prediction result and in response competitors will be awarded with public score based on their submission until the deadline of the competition finishes, but public score is being resolved from an only fraction of the test data set I between(25-33%). Each participant can improve their performance based on public Leaderboard score results.

Private Leaderboard

When the competition ends, kaggle selects best submission result among all your submissions and score your prediction against the "remaining fraction" of the test

[8]https://www.kaggle.com/solutions/competitions

set. Competitors never receive ongoing feedback about individuals or team score, respectively because of that it is private Leaderboard. So final competition results are based on private Leaderboard only.

4.8 Kaggle : Data Science Job Board

Hiring Data scientist

Access 2,76,325 Data Scientist. The Kaggle is the world's largest community of data scientists and machine learning engineers. Kagglers users demonstrate the skills to solve the toughest problems across many industries. Kaggle not only host the competitions but also it helps companies to hire the data scientists from their website which gives encouragement to the kaggler to do better in the competitions and get hired by the one of the top leading companies or started new companies.

Seeking:

browse 869 careers kaggle job board sources, career opening as a data professionals. Once a user subscribed Kaggle will notify the user with the new opportunities in the data science, machine learning, Statistics and other analytics jobs.

4.9 How Kaggle has changed the life of Data scientist

Figure 9: Jeremy Howard[9]

Figure 10: Owen Zhang[10]

Jeremy Howard (figure:9) started his career as a management consultant at some of the world's most exclusive firms such as wunderkind, McKinsey & Company, AT Kearney. In 2010 Howard participated in Kaggle competitions and ranked top most participants in data science competitions in both 2010 and 2011.After participating in Kaggle his life has been changed completely he was a president of Kaggle and promoted Kaggle now at present his own a company which is started in August 2014 named as "Enlitic" with the mission of leveraging recent advances in machine learning.

In past Owen Zhang (figure: 10) worked as Travelers Insurance and he started participating in Kaggle competition got 1st ranked and earned 774,329 points by competing in 28 competitions (2013-15) and later he got hired by "American International Group". At present, Zhang working as a Chief product officer at "Data Robot".

5 My experience in Kaggle competition

I have experimented myself by participating in one of the challenges posted on Kaggle website. The challenge of this data science project is to classify handwritten digits using MNIST data, in simple, its called as Digit Recognizer. The main goal of this competition is to introduce the people with machine learning concepts. This competition is to take an image of handwritten single digit and determine what digit it is. The data for this competition is taken from the MNIST dataset (Modified National Institute of Standards and Technology). The data file consists of two data sets, training and testing, respectively which contains grayscale images of hand-drawn digits from one to nine. Each image is 28 pixels in height and 28 pixels in width, for the total of 784 pixels in total and each pixel has a single pixel-value associated with it which has light and dark colors where dark color indicates the higher value of the numbers. The pixel value ranges between 0 and

255. The data set of training set has 785 columns in total where the first column is called a label, whereas the rest of the columns contain the pixel-values of the associated image. Every pixel column in the training set has a name like pixel x, where x is an integer between 0 and 783. Consider an example of the image having a decomposition of x such as x = I * 28 * j, where I and j are integers between 0 and 27, then it position is located on the row of I and column j of 28 * 28 matrices. From the ASCII-diagram, we can identify the position of the images. If we omit the pixel prefix, the pixels make up the image as shown below.
000 001 002 003 ... 026 027
028 029 030 031 ... 054 055
056 057 058 059 ... 082 083

| | | | ... | |

728 729 730 731 ... 754 755
756 757 758 759 ... 782 783 The testing data set is same as the training dataset, but does not contain the column label as we have to identify the number and put on labels for the output. The expected output is supposed to be as explained in the below.

```
3
7
8
(27997 more lines)
```

The dataset has a label for the digit and 28 * 28 = 784 columns of values at each pixel at the intersection. There are many ways to solve this problem. We can use KNN models, SVM, random forest, SVM etc. Here we explain in detail about training via KNN models.

How does Random Forest work?

Random Forest is a trademark term for an ensemble of decision trees. Unlike single decision trees which are likely to suffer from high Variance or high [Bias]. It uses averaging to find a natural balance between the two extremes.

Since they have very few parameters to tune and can be used quite efficiently with default parameter settings (i.e. They are effectively non-parametric) Random

Forests is best to use as a first cut when we don't know the underlying model, or when it is required to produce a decent model under severe time pressure.

Random Forests an ideal tool for people without a background in statistics, allowing people to produce fairly strong predictions, free from many common mistakes, with only a small amount of research and programming.

digit Recognizer code:

```
library(devtools)
install_github("hadley/readr")
install_github('dmlc/xgboost',subdir='R-package')
library(randomForest)
library(readr)

set.seed(0)

numTrainset <- 10000
numTrees <- 25

trainset<- read_csv("train.csv")
testset <- read_csv("test.csv")

rows <- sample(1:nrow(trainset), numTrainset)
labels <- as.factor(trainset[rows,1])
trainset <- trainset[rows,-1]

randfn <- randomForest(trainset, labels, xtest=testset, ntree=numTrees)
predictions <- data.frame(ImageId=1:nrow(testset), Label=levels(labels)[randfn$test
head(predictions)

write_csv(predictions, "randfn_benchmark.csv")
```

In the above code, random forest algorithm will create a multiple random subset. Every random subset will create a decision tree and every decision will have different variance and all this generated decision will be again used for ranking of the

classifier. Later, the results are exported to .csv file and the Output is shown in the figure 11. Output is exported in such a way that the output is written in a single line with digit which is predicted. The output shows that the first image is predicted as 2 and the second image is predicted as 0, the third image is predicted as 9 and so on.

Output

A	B
ImageId	Label
1	2
2	0
3	9
4	4
5	2
6	7
7	0
8	3
9	0
10	3

Figure 11: output predictions results in csv format[11]

Public Leaderboard

The above figure: 12 shows the position in Leaderboard once I submitted my prediction results in kaggle in fraction of seconds I can able to view my rank, i.e. 364 with the accuracy rate of 93%.figure shows the Public Leaderboard of Kaggle.

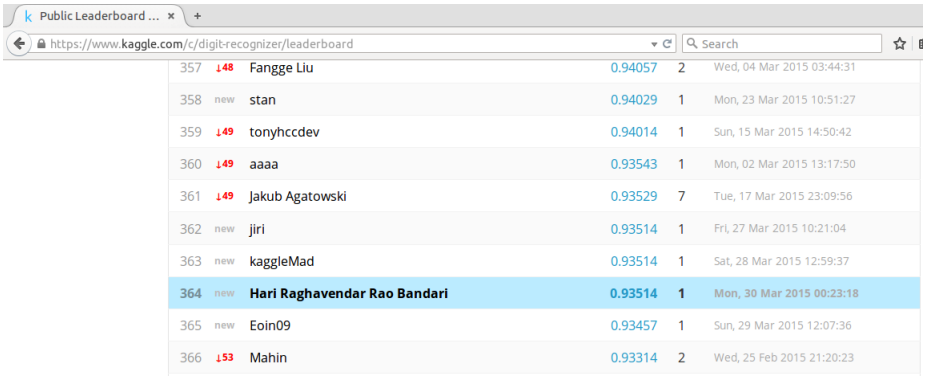

Figure 12: Kaggle Public Leaderboard[12]

6 Recommendation

My experience after participating in kaggle competition digit recognizer, I read and explored about "Introduction to R". It was very easy and challenging to participate in the competition and gave much knowledge how kaggle website works. In near future I will participate in many competitions to test my skills and to improve my technical skills.

People who participate in competitions will be hired by many top companies as data scientist or data analyst, etc. Because participants will explore real world prediction problems and apply different techniques to solve it. As an example, I explained about two data, scientist's life changed after participating in kaggle. In addition to the prize money and data, they use Kaggle to meet, learn, network and collaborate with experts from related fields.

kaggle host competitions in well structured way and well organized to work in teams.

7 Conclusion

In this report, Kaggle is a platform for hosting public data science challenges, in which sponsors post their complex problem to the Kaggle platform, then data scientists from all over the world compete in the respective challenge to create the best solution for the data problem. Another advantage of competing in the kaggle competition is to get hired by companies as a data scientist based on the points where each individual obtained by solving the data science problem. In conclusion, Kaggle is a platform for making a data science as a sport because it always host the data science related unsolved complicated problems, then every data scientist will compete in the competition and will be very active and it will be very challenging as public Leaderboard will be always updating based on the proposed solution accuracy of the data problem.

References

[1] Harvard Business Review, *Data Scientist: The Sexiest Job of the 21st Century*. by Thomas H. Davenport and D.J. Patil ,OCTOBER 2012, url = https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[2] McKinsey & Company, *Big data: The next frontier for innovation, competition, and productivity*. by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers ,May 2011.

[3] Pivotel, *McKinsey Report Highlights the Impending Data Scientist Shortage*. by Paul M. Davis ,july 23 2013. url = http://blog.pivotal.io/pivotal/news-2/mckinsey-report-highlights-the-impending-data-scientist-shortage

[4] CitizenNet, *Three Trends in Marketing Technology for 2014*. by Dan Benyamin ,july 3 2014. url = https://citizennet.com/blog/2014/01/03/four-trends-in-marketing-technology-for-2014/

[5] NETFLIX, *Netflix Prize Completed*. september 21,2009 url = http://www.netflixprize.com/index

[6] MASTER'S IN DATA SCIENCE, *Data Scientist Foundations: The Hard and Human Skills You Need*. by DS Examiner ,November 8, 2013. url = http://www.mastersindatascience.org/blog/data-scientist-foundations-the-hard-and-human-skills-you-need/

[7] WIRED, *Machine Intelligence Cracks Genetic Controls*. by Emily Singer, Quanta Magazine Science ,December 29, 2014. url = http://www.wired.com/2014/12/machine-intelligence-cracks-genetic-controls/

[8] Rstudio, url = http://www.rstudio.com/products/rstudio/download/ url = http://www.statmethods.net/

[9] Kaggle, Digit Recognizer url = https://www.kaggle.com/c/digit-recognizer