IN LEARNING YOU WILL TEACH, IN TEACHING YOU WILL LEARN

Data Science with R (DSWR) is one of the largest groups on Facebook for data science, Machine Learning, R, Python. If you are a newbie, scroll through this booklet to have a clear understanding.

Etiquettes

Posting is strict at DSWR. Please realize that we are here to learn from each other, thus spamming is neither desirable not appreciated.

Here what we delete immediately, and you may get banned for violating these rules

- Re-post from a page,
  Re-post of an article which is already posted
- Articles that are directly not related to data science, R etc
- Copyrighted material, unless the author explicitly releases the content
- Brand promotion: articles/PR posts typically mentioning with your company's name and brand counted as promotion. Violating this rule might lead to your immediate ban.
- Questions which can be easily answered by google search.
- Questions which have been already answered in the FAQs

Topics covered in this booklet

Data Science With R- FB Group

Data Science With R- FB Group

How to Start with R and Data Science

- Introduction to Statistical Learning with R - A go to book to start data science journey using R (freely available)
- Statisical Learning in 15 Hours (Freely Available; by Trevor Hastie and Others. Most recommended course on ML with R)
- Analytics Edge on EDX (Arguably one of the best introductory courses on ML with R)
- Predicting Titanic survivors With R - A guided tutorial on how to do your first data science Kaggle project
- R Bloggers tutorial series for learning R - A master compilation of links from a network of R enthusiasts
- R For Data Science - An online book with a guided tour through the process of doing data science; written by Hadley Wickham
- Other resources: MOOCs such as Edx, Coursera, Udacity, Udemy, datacamp & Youtube have plenty of courses to start with

Statistics for Data Science (Required study and resources):-

- Linear algebra-Linear Algebra– MIT 18.06 Linear Algebra by Gilbert Strang
- Probability theory-Probability and Statistics– MIT 6.041 Probabilistic Systems Analysis and Applied Probability by John Tsitsiklis
- Calculus
- Multivariate Calculus
- Graph theory
- Optimization methods
- Elements of Statistical Learning (For in-depth statistical derivation of many ML and statistical techniques)
- Introduction To Statistics Introductory course from Variance Explained
- Other useful Sites:- Youtube, Khan Academy, Google Search

Open Source Data Repository

- R library ISLR has following 15 datasets. You can access them by typing "*library(ISLR)*" and "*attach(USArrests)*"

1. **Auto** Gas mileage, horsepower, and other information for cars.
2. **Boston** Housing values and other information about Boston suburbs.
3. **Caravan** Information about individuals offered caravan insurance.
4. **Carseats** Information about car seat sales in 400 stores.
5. **College** Demographic characteristics, tuition, and more for USA colleges.
6. **Default** Customer default records for a credit card company.
7. **Hitters** Records and salaries for baseball players.
8. **Khan** Gene expression measurements for four cancer types.
9. **NCI60** Gene expression measurements for 64 cancer cell lines.

10. **OJ** Sales information for Citrus Hill and Minute Maid orange juice.
11. **Portfolio** Past values of financial assets, for use in portfolio allocation.
12. **Smarket** Daily percentage returns for S&P 500 over a 5-year period.
13. **USArrests** Crime statistics per 100,000 residents in 50 states of USA.
14. **Wage** Income survey data for males in central Atlantic region of USA.
15. **Weekly** 1,089 weekly stock market returns for 21 years.

- IRIS dataset- available in datasets function of R
- UCI ML Datasets- 452 different datasets from various fields
- Other sites- Kaggle, Google Dataset Search

Advanced R With ML -

- Advance R Programming - Coursera by Roger Peng
- Elements of Statistical Learning (Freely available by Tibshiarani et al)
- ggplot2: Elegant Graphics for Data Analysis (Use R!)- Book by Hedley Wikham (Amazon Link)
- Advanced R (Freely accessible online; by Hall & Chapman)
- R Packages -
- Text Mining With R (Freely accessible online; by Julia Silge and David Robinson)
- Introduction to Spatial Analysis with R

FAQs

FAQs has answers to the following question (Please note, answers here largely reflect the views of the author of this document. You are welcome disagree with one or more points).

A. **Which is good to learn R or Python**? Hard to say. Both are good and widely used in data science; R is for quick ML/Statistical analysis while Python is used for production/deployment; generally speaking. R can be used for production as well but it aint as much scalable as Python in my opinion, others might disagree.
Most importantly, It's always better two know two programming languages than one. However, it is always better to be awesome at one thing than sucking at two things simultaneously.

B. **I have a dataset but no idea which ML algorithm to use**? If your response variable is continuous, you can use linear regression, KNN, regression trees (CART, RF, GBM, XGB). If it's categorical variables then use Logistic regression (works quite well for binary categories), Classification trees, SVM, KNN etc. These are the basic ones. There are so many advance other algorithms for decomposition, regularization, dimensionality reduction, family of neural nets etc.

Having said that, there are no clear cut answers for this question. Over a period of time, you would develop intuition which algo works best for what kinda dataset. Data pre-processing steps are way more important than using any algorithm.

Data Science With R- FB Group

C. **Is it important to have Maths/Stats background to get into data science**? Not really. But again you need to have fair understanding of Linear Algebra, Calculus, Matrix Theory, Maximum likelihood, Proability, Bayes theorem, Graph theory etc.
If you dont like to study at maths/Stats, you are strongly should reconsider data science as a career.

D. **What are some important R packages?** These are good to know R packages (not in any particular order) caret, caTools, dplyr, randomForest, xgb, InformationValue, mice, readr, ISLR, MASS, NLP, SnowballC, tm, ggplot2, lubridate etc

E. **I have a dataset with missing values. What should I do**? You can do one or more of the following

1) You can replace missing values with mean, mode, median; depending upon the variable type
2) You can remove those variables which have 20% or more missing data.
3) You can use randomforest, linear regression, logit regression to predict missing values (depending upon whether a variable is category or numeric)
4) You can check the values are missing at random (MAR) and can use mice (multivariate imputations of chained equation) algorithm to fill missing values accordingly
5) You can create bins; check the information values of missing data to figure out whether missing values are really important.

F. **Is outlier treatment necessary before training my model?** Generally yes. If you are working on a model which includes some kind of distance (e.g. linear & logistic regression, SVM, K-Means etc), then outlier treatment is necessary. It's not much important for trees algorithms like RF, CART etc

G. **What are some good clustering algorithms with their pros and cons?** There are three major clustering algorithms

1) K- Means- Extremely fast and easy to use. Occasionally suffers with random initialization trap
2) Hierarchical Agglomerative- Gives you nice dendrogram to decide how many clusters to use. Computationally expensive if the dataset is huge
3) DBSCAN- Robust to noise. Sensitive to clustering parameters

H. **How do I deal with multicollinearity in my dataset?** There are couple of way to deal with multi-collinearity

1) Remove the variables that have high correlations
2) Use stepwise regression. It will give you an optimal model at the end
3) Use regularization like Ridge or Lasso in the model training to reduce impact of multicollinearity
4) Use Factor Analysis or PCA and convert your data into low dimensions and use these dimensions as your new features

I. **I have a dataset with huge no. of features. How can I reduce the dimensionality of the data?** There are four methods for this

Data Science With R- FB Group

1) **Univariate Selection**:- Here we look at anova/t-test/correlation/chi-square test of an independent feature with target variable and decide whether it should be kept in the model
2) **Variable Selection from Model**:- Some model like Linear Reg (p-value), Random Forest (Variable importance) give some kind of information about the importance of variables in the model. We can select variables accordingly and retrain our model
3) **PCA**- Convert your data into low dimensions using PCA and use these components as your new transformed features
4) **Iterative Selection**:- Here we use stepwise regression/classification (forward, backward or hybrid model)

J. **How to work on text data which has sentences as one or more features?** For text data, you can use Bag-Of-Words method to extract important words and use them as features. Then you can do feature-scaling using tf-idf or use n-Grams

K. **How to determine which model is better?** You can look at test set accuracy parameters. For e.g. if it's linear regession, the model with highest adjusted R-squared, lowest MSE, RMSE, RMLSE, AIC etc would be better in general. It it's logistic regression, the model with higher accuracy, higher auc & ks statistic would be better. In general, a model with comparatively higher accuracy on test set and lower misclassification rate would be considered better.

L. **How to deploy models from scratch?** There are couple of ways.

1) You can create a scheduler (in Windows) or CRON (in linux) which would run the script after a predetermined interval.
2) In ML Azure Studio you can write R script and upload pre-trained R models
3) If your dataset is huge, build a pipeline using SparkR & R
4) You can use AWS EC2 and Shiny App as well.
5) If you are looking for full scale deployment of your ML models, Python is widely recommended

M. **My dataset doesn't meet the linear regression assumptions. What to do now?** If your feature is numeric and doesn't follow normal distribution, you can use one of the following methods

1) If the variable is positively skewed, use its log, square root & cube root. If data has zero take Log(X+1) instead of just Log(X) because log0 is undefined
2) For left-skewed data—tail is on the left, negative skew—, common transformations include square root (constant – x), cube root (constant – x), and log (constant – x).
3) You can also use general power transformation, such as Tukey's Ladder of Powers or a Box–Cox transformation. These determine a lambda value, which is used as the power coefficient to transform values. X.new = X ^ lambda for Tukey, and X.new = (X ^ lambda – 1) / lambda for Box–Cox.

Data Science With R- FB Group

N. **I have a date column, how do I use date as feature(s)?** You can use strptime() function or use lubridate to extract, day, month, year, hour etc from the date column and use em as variables.

O. **Which Laptop/Machine should I get for data science?** If you plan to do data science (excluding deep learning), any machine with 16+GB RAM, i7 processor with Ubuntu/Windows and 1TB HDD would do fine. If you plan to extensively use deep learning, you may get an assembled machine with 32/64 gig RAM, GPU and Ubuntu. Use google to get more info on this.

P. **What are some good podcasts for datascience?** There are so many of them. You can google them. These are I personally find interesting:- Machine Learning Guide, Linear Digressions, Data Science at Home, Half Stack Data Science, Data Framed, Talking Machines, Learning Machines 101, All Things Data, Bonus Feed, SuperDataScience etc

Admin's Note
While I understand I can't answer possibly all FAQs, I have tried my best with this small booklet to cover as much as I could. Hope, this small effort would help our new members having a better clarity on data science. If you have any query/comment, please write it in the comment box. I will try to answer it ASAP.

IN LEARNING YOU WILL TEACH, IN TEACHING YOU WILL LEARN

Signed,
Sohail Ahmad
23RD Oct'18

Data Science With R- FB Group