WORK REPORT

ORGANIZATION: DecisionStats
DURATION: 1st June, 2015 - 11th July, 2015
TYPE: Data Scientist
TYPE: Internship

CONTENT:

PART 1: PRE INTERNSHIP ACTIVITIES:

After being selected as a data science intern at DecisionStats, I had to complete a set of tasks before joining. They were:

1. Learning basics of Python- I achieved this by taking part in the Spoken-tutorial by IIT-Bombay, and earned a score of 95%.
   COURSE LINK- http://python.fossee.in/
2. Completing Beginners course at DataCamp- This course was a 4 hours introduction to concepts in R and worked as a refresher course for me.
   COURSE LINK- https://www.datacamp.com/courses/free-introduction-to-r
3. Installation and hands with various tools for R and Python which included- Anaconda packages, Ipython, Ipython notebook, Rstudio, spyder and IDLE. Of all the tools, I mostly worked on ipython notebook and Rstudio.

PART 2: INTRODUCTION TO BLOGGING:

As a part of our internship we had to set up and maintain a blog. The blog would hold the snippets of what we are learning. It was through blogging that I learned how to be more presentable with my work. With Mr. Ajay and Miss Sunakshi as role models, people who have blog hits in 20k+, I learned about how to optimise my blog with title, codes, tags and categories.
MY BLOG- https://themessier.wordpress.com

PART 3: PANDAS IN PYTHON:

Pandas is a very powerful and fun to use library in Python, supporting data set from csv, excel, json and mysql in a single package. Pandas offer easy handling of data frames, missing values, slicing and grouping, that helps dissect huge dataset and extract useful statistical information and correlations.

Learning Pandas led to learning features of numpy (an exhaustive mathematical library in Python) and matplotlib(A beginner level plotting library in Python)
COURSE LINK- https://bitbucket.org/hrojas/learn-pandas

PART 4: RCPP using R AND C++:

R is powerful but can be slow, C++ is fast but not data oriented, by converting a part of R code in C++ code we can have faster results. Rcpp does the same. By writing codes in R and C++ and comparing their time, I learned how to write C++ equivalent of R and incorporate in my R code.
COURSE LINK- http://adv-r.had.co.nz/Rcpp.html

PART 5: SUMMER SCHOOL 2015:

As an intern at DecisionStats I got a chance to attend the Summer School in Data Analytic, for free. The 10 days course was divided in 4 part.

1. **Introduction to Analytic**- Starting with basic definition of data science, hacker, logical fallacies and various business models, I learned about business intelligence, customer evaluation using the LTV and RFM analysis.
2. **Business Analytic in Python** - Detouring through basics in Python, the pandas and numpy libraries, the highlight of the day was data visualization using the matplotlib, ggplot, KDE plots.
3. **Business Analytic in R**- The most awaited session, it taught me about using geodata vi ggmap, using GUI based libraries like rattle and deducers to do in-depth statistical analysis and modelling, the data visualization section introduced us to high end libraries like brewer and lattice.
4. **Introduction to SAS**- I worked on SAS university edition and gained hands on basic SAS, the data command and proc commands of reading, presenting and performing calculations on the data.
   COURSE LINK- https://github.com/Decision-Stats/ppts

PART 6: TRAINING CODES:

This involved creating the decisionstats organization on github and adding the presentations and codes used during the training. Apart from acting as great revision session, it helped me in learning the functions at my pace. It taught me the importance of maintaining proper code for each project one undertakes, and open sourcing it for others to learn from.
I came across nbviewer and R pubs to share my Ipython and RStudio codes.
CODE LINK- https://github.com/Decision-Stats/s15_codes

PART 7: WEBSITE:

Done as a side project, I along with 2 other interns, helped in re creating the website of DecisionStats. Although, static and done with minimal HTML/CSS, it was the first time we worked on a professional site and employed aesthetics and design thinking.

SITE LINK- http://DecisionStats
CODE LINK- https://github.com/sara-02/decision_site

PART 8: FINAL PRESENTATION:

I worked extensively on ggplots in Python and as a part of the final assessment had to prepare a presentation about the same, it involved setting up ggplot, going through the documentation (which helped me in reporting few errors in their documentations), and implementing each function on my ipython. To make the presentation easy to grasp I added side notes for each new function. The overall review will come after I present it to the team in the final meeting.

PRESENTATION LINK- http://tinyurl.com/ggplot-Sarah

PART 9: FURTHER LEARNING:

What I learned during a month was an introduction to data science, there is an exhaustive list of tools which we will learn in the coming months, few of them are:

- R in cloud computing- Finished the basic AWS setup for Rstudio.
- Hadoop and Mapreduce
- Swirl in R
- Shinly in R
- Applying rcpp on real world problem.
- Machine learning and astronomy packages in Python.
- Modelling and classification on real world dataset.
- Basic hacking on Kali Linux
- Entrepreneurship 101/102 on edx
- Data science specialization from Johns Hopkins.
- Kaggle projects.

CONCLUSION:

Working in an start up has its own pros and cons, one does not get a cosy workplace but apart from gaining technical knowledge, one gets hands on what to do and what not to do while running a startup. I gained a lot on personal and professional end by working with a diverse team of learners and industry experts and hope to keep up with my progress even after the completion of my internship.

I, hereby declare that the information given above is true to the best of my knowledge.

-SARAH.

Miss Sarah Masud,
Intern,
DecisionStats.

Mr. Ajay Ohri,
Founder,
DecisionStats.