Analysis of data from the survey with developers on Stack Overflow: A Case Study

[1]Y. Beeharry; [2]M. Ganoo
[1]Department of Electrical and Electronic Engineering, Faculty of Engineering
University of Mauritius
Réduit
Mauritius

[2]Faculty of Information, Communication, and Digital Technologies
University of Mauritius
Réduit
Mauritius
y.beeharry@uom.ac.mu, manish.ganoo@gmail.com

Abstract*: *Many businesses are understanding the current evolution of Big Data Analytics around the world. Along this line, businesses are investing enormously in view not to lose competitive advantage. The work in this paper, analyses the data from the survey conducted with the numerous developers on Stack Overflow in order to gain insights on the directions of programming languages, databases, and the job seeking status of the developers. Results show that the trend is developers want to use more the programming languages and databases used on cloud platforms for Big Data Analytics. Additionally, a Distributed Random Forest model with 87.64% accuracy, for predicting the job seeking status of developers shows that the developers may not be looking to move to new job environments and would prefer staying in their current company or organization. This would be an indication that developers are most probably looking forward to bring added value to their current companies/organizations as Big Data Analytics would start to be adopted.

***Keywords*: Programming Languages, Databases, Software Developers Survey, Distributed Random Forest.**

(Article history: Received: 24th June 2018 and accepted 22nd October 2018)

I. Introduction

Data science and analytics is the most booming field of the present era [1]. All businesses are looking for expertise in tapping information that they generate without using to a great extent as well as for extracting valuable insights in order to gain competitive advantages and increase profits. International Business Machines (IBM) which is one of the leading Information Technology and cloud service provider forecasts a rise of about 28% in the demands for data scientists by the year 2020 [2]. The study by IBM also suggests that the industries where data science jobs with big data and machine learning skills would be in highest demands are Information Technology, Finance and Insurance. The soar in sensor data in the field of Internet of Things (IoT) has engendered a state of data inflation which would also be requiring data science and analytics expertise for obtaining insights and performing predictions [3].

The primary focus for becoming a data scientist is to use technology for the examination of large volumes of raw data and develop predictive models as well as obtain insights for large target audiences [4]. Along this path, the data scientist has to acquire cross-disciplinary skills which is a very scarce quality. Thus, the major model for properly operating data scientists is a team consisting of members with cross-disciplinary skills. These skills are mainly: Mathematics, Statistics, Data Storage / Warehousing, Programming, among others. One of the major challenges is the requirement for programming skills in order to succeed as a data scientist because leaders in the tools of big data science develop different tools and technologies to be used with different programming languages for statistical analysis [5].

A report from Gartner Inc. brings forth the prediction whereby there would be automation in more than 40% of the data science tasks by the year 2020 [6]. The report on the magic quadrant for data science platforms puts IBM, Statistical Analysis Systems (SAS), Rapid Miner, and Knime as leaders for the year 2017 with new inclusions compared to the quadrant of 2016. Some of the interesting new inclusions in the quadrant for the year 2017 are: Math works as challenger, H2O,ai as Visionaries, and Teradata as Niche Players. Among the leaders, IBM has a more diverse panoply for data analytics in terms of technologies and programming languages to be used [7]. The most considered programming languages in the field of data science and analytics are: R, Java, Python, Scala, JavaScript, and Clojure [5].

Some research has been performed on First Programming Languages (FPL) and reproducible research analyses based on programming languages. For example, in [8], the problem of the variation in the pool for FPL at different times has been analysed. The authors have proposed a framework for the evaluation and comparison of different object-oriented programming languages to assess their suitability for the pool of FPL. Additionally, in [9], the research reproducibility issue across the world has been addressed. The major known cause of this hurdle is the fact of having researchers with knowledge of dissimilar programming languages. In this case, a major biomedical project has been considered and the researchers have developed a reproducible computing tool which allow a tool-agnostic approach to biomedical data analysis. A combination of relational database, statistical computing environment, and standard command-line tools has been used for the development of the tool.

Predicting the rise in demand for data science and analytics job prospects requires to be backed up by some analysis on the current world-wide community of programmers in terms of the programming languages they currently use, the computing systems they are currently familiar with, their education level and job satisfaction in the field, willingness to learn new technologies, in addition to other critical and insightful aspects. These analyses can be beneficial by providing knowledge for the way ahead to having major contributions in the field of data science and analytics throughout the world.

The work in this survey paper is based on the analysis of the survey data filled by developers on stack-overflow around the world. Section 2 provides some details on the survey data used and analyses to be performed. Section 3 provides the analyses and results obtained and the work is finally concluded in Section 4 with some future works.

II. Survey Details

The dataset used for analysis in this work is the Stack overflow survey (2017) dataset available from Kaggle [10]. The initial dataset consist of more than 64,000 responses. The data munging process in this work has narrowed down the dataset to a balanced one with 525 male and 525 female respondents.

The main analysis performed in this work are:

- Programming languages that developers have used v/s Programming languages that developers want to use.

- Databases that developers have used v/s Databases that developers want to use.

- Distributed Random Forest based classification algorithm for prediction of the job seeking status of respondents.

The classification algorithm used in this work is the Distributed Random Forest model [11, 12]. The JobSeekingStatus of the developers has been modelled using different parameters in the dataset. The parameters in order of importance for the Distributed Random Forest model is as shown in Figure 1. This order of importance has been used for the parameter selection of the model.

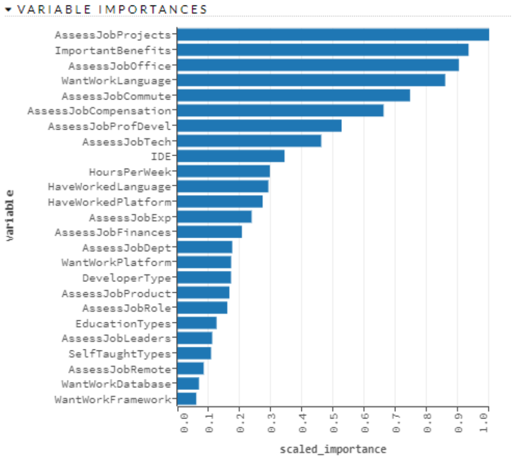


Fig. 1. Importance of Parameters in the Distributed Random Forest model

The details on the different variables are shown in Table 1

TABLE I. Table Type Styles

Variable	Representation
JobSeekingStatus	Description of current job-seeking status.
AssessJobProjects	Assessment of potential jobs to apply from the perspective of project management at the company or organisation.
ImportantBenefits	Most important compensation and benefits, other than base salary.
AssessJobOffice	Assessment of potential jobs to apply from the perspective of office environment to work in.
WantWorkLanguage	Languages with extensive development work in over the past year, and which do you want to work in over the next year?
AssessJobCommute	Assessment of potential jobs to apply from the perspective of amount of time to be spent commuting.
AssessJobCompensation	Assessment of potential jobs to apply from the perspective of compensation and benefits offered.
AssessJobProfDevel	Assessment of potential jobs to apply from the perspective of opportunities and professional development.
AssessJobTech	Assessment of potential jobs to apply from the perspective of languages, frameworks, and other technologies willing to be working with.
IDE	Development environments used regularly.
HoursPerWeek	Number of hours spent on activities related to finding new job opportunities.
HaveWorkedLanguage	Languages with extensive development work done in over the past year, and which do you want to work in over the next year?
HaveWorkedPlatform	Platforms with extensive development work done in over the past year, and which do you want to work in over the next year?
AssessJobExp	Assessment of potential jobs to apply from the perspective of experience level called for in the job description.
AssessJobFinances	Assessment of potential jobs to apply from the perspective of financial performance or funding status of the company or organization.
AssessJobDept	Assessment of potential jobs to apply from the perspective of specific department or team to be working on.
WantWorkPlatform	Platforms with extensive development work done in over the past year, and

Variable	Representation
	which do you want to work in over the next year?
DeveloperType	Best description of the type of developer.
AssessJobProduct	Assessment of potential jobs to apply from the perspective of how widely used or impactful the product or service to be working on is.
AssessJobRole	Assessment of potential jobs to apply from the perspective of specific role and job title.
EducationTypes	Education outside the formal schooling.
AssessJobLeaders	Assessment of potential jobs to apply from the perspective of reputations of the company's senior leaders.
SelfTaughtTypes	Resources used for self-taught programming technology without taking a course.
AssessJobRemote	Assessment of potential jobs to apply from the perspective of the opportunity to work from home/remotely.
WantWorkDatabase	Database technologies with extensive development work done in over the past year, and which do you want to work in over the next year?
WantWorkFramework	Libraries, frameworks, and tools worked with extensively over the past year, and which do you want to work in over the next year?

The Job Seeking Status can be classified in the following three classes:

Class 1: Actively looking for a job,

Class 2: Not interested in new job opportunities, and

Class 3: Not actively looking, but open to new opportunities.

The different results obtained are given in the following section.

III. Results and Analysis

The analyses performed in this paper are: finding correlation between the programming languages, databases (NoSQL) and job seeking status of respondents. The graph representing the number of developers who want to work with specific programming languages versus the number of developers who have worked with specific programming languages is shown in Figure 1.

Figure 2 shows the graph representation of the number of developers who want to work with specific databases versus the number of developers who have worked with specific databases. It can be observed that most developers have used structured databases like: MySQL, SQL Server, and PostgreSQL; and want to continue using them. However, it can also be observed that few developers have used NoSQL databases like MongoDB and more developers want to use them as compared to the structured databases. This gives a good indication for the trend in NoSQL databases being used for big data analytics on cloud platforms.

It can be observed from Figure 3 that most developers code in JavaScript, SQL, Python, Java, and C#. This gives a good indication for the trend in programming languages being used for big data analytics on cloud platforms [13]. With the data collected from the survey, a distributed random forest model has been used for modelling the job seeking status of all the developers using the most significant variables as depicted in Section 2. H2O has been used as the open-source tool to build the model and test its accuracy. The data has been split into 75% and 25% for the training dataset and test dataset respectively. Figure 4 shows the Distributed Random Forest model construction snapshot.

The model is then tested using the test dataset. Figure 5 shows the prediction outputs of the model on the test dataset. It can be observed that a very low Mean Squared Error (MSE) of 0.0971 is obtained. The r-squared value is 0.783 which gives an indication of the correlation of the model. The correlation value being closer to 1 demonstrates that the model can predict correctly around 78% of the time.

Figure 6 shows the confusion matrix of the prediction performed on the test dataset. It can be observed that with the Distributed Random Forest model, Class 1 is predicted with 100% error, while Classes 2 and 3 are predicted with greater than 99% accuracy. This shows that when the Distributed Random Forest model implemented in this work is used, the Classes 2 and 3 are predicted more accurately than Class 1.

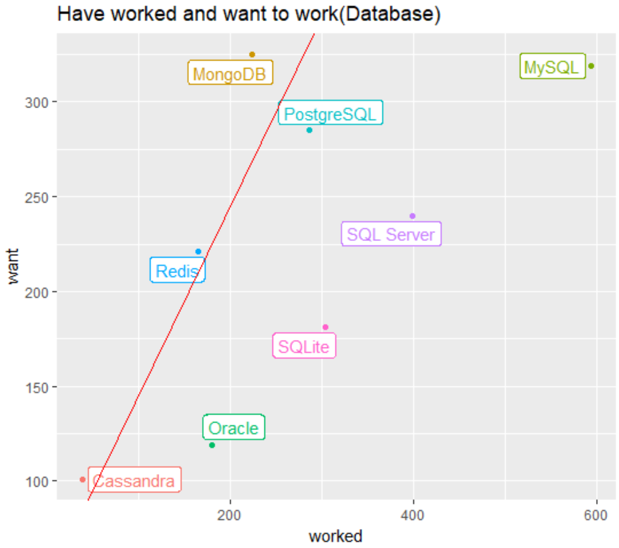


Fig. 2. Number of developers who want to work with specific databases versus the number of developers who have worked with specific databases

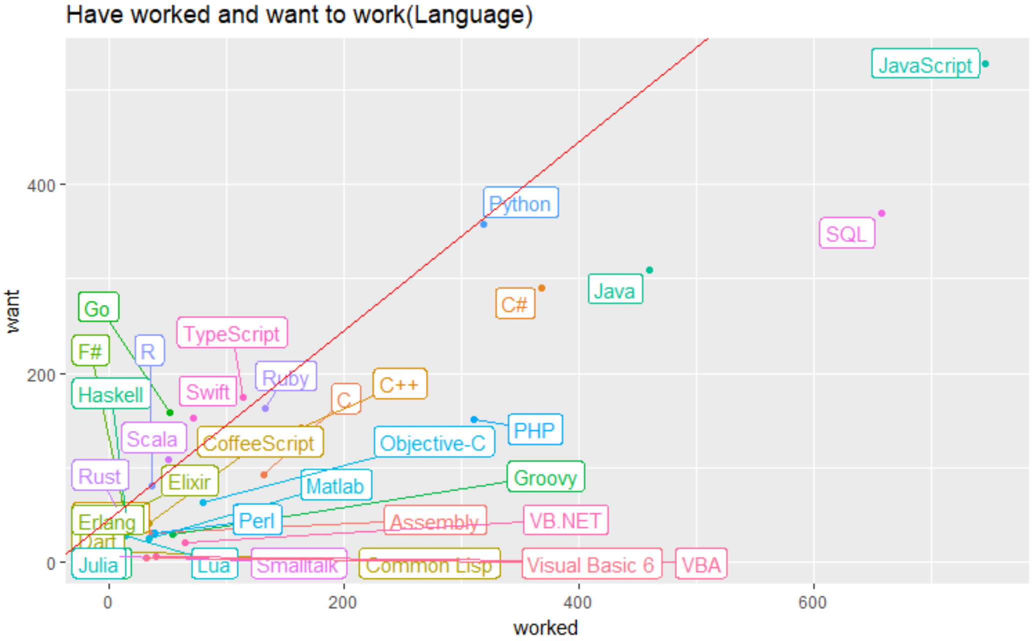
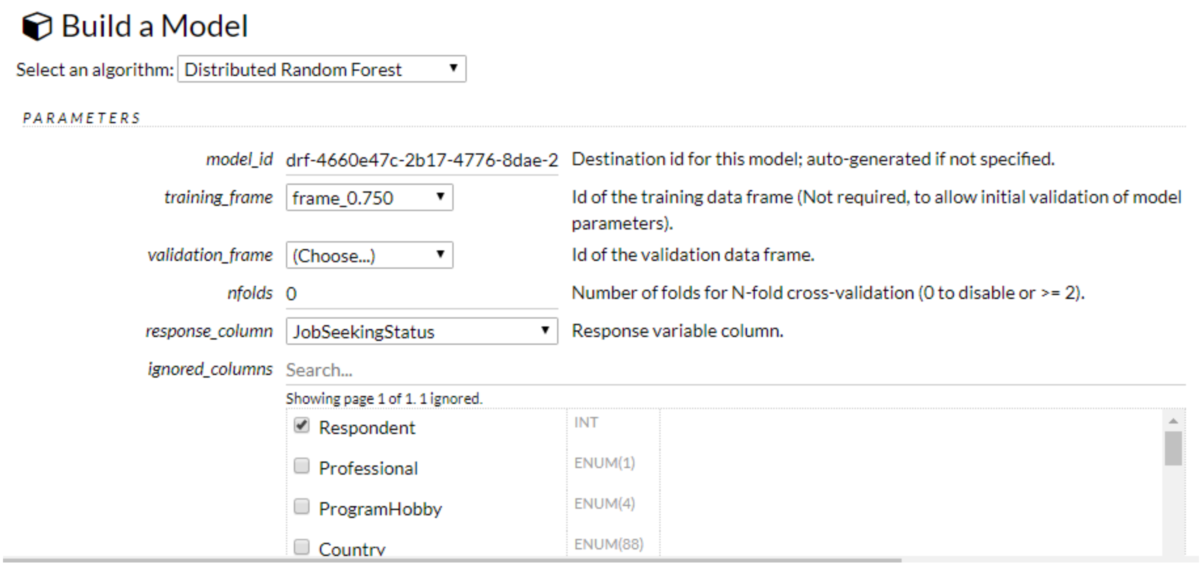


Fig. 3. Number of developers who want to work with specific programming languages versus the number of developers who have worked with specific programming languages



Build a Model

Select an algorithm: Distributed Random Forest ▾

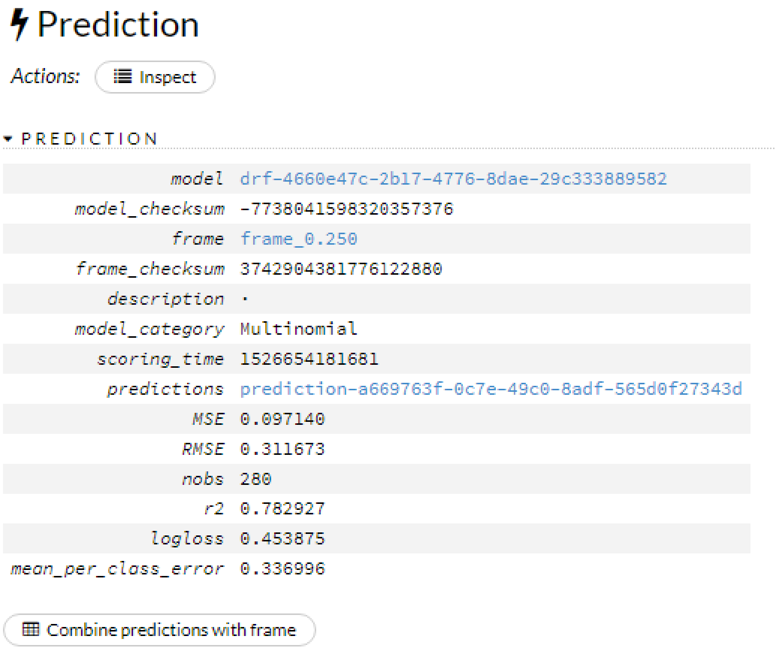
PARAMETERS

model_id	drf-4660e47c-2b17-4776-8dae-2	Destination id for this model; auto-generated if not specified.
training_frame	frame_0.750 ▾	Id of the training data frame (Not required, to allow initial validation of model parameters).
validation_frame	(Choose...) ▾	Id of the validation data frame.
nfolds	0	Number of folds for N-fold cross-validation (0 to disable or >= 2).
response_column	JobSeekingStatus ▾	Response variable column.
ignored_columns	Search...	

Showing page 1 of 1. 1 ignored.

☑ Respondent	INT
□ Professional	ENUM(1)
□ ProgramHobby	ENUM(4)
□ Country	ENUM(88)

Fig. 4. Distributed Random Forest Model with 75% of the dataset



Prediction

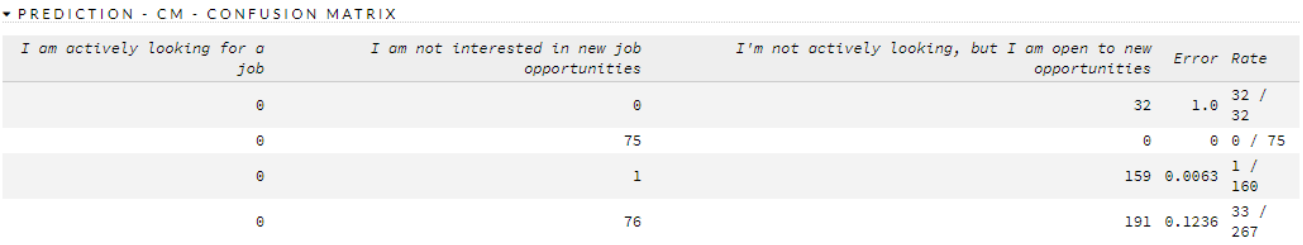
Actions: Inspect

PREDICTION

model	drf-4660e47c-2b17-4776-8dae-29c333889582
model_checksum	-7738041598320357376
frame	frame_0.250
frame_checksum	3742904381776122880
description	·
model_category	Multinomial
scoring_time	1526654181681
predictions	prediction-a669763f-0c7e-49c0-8adf-565d0f27343d
MSE	0.097140
RMSE	0.311673
nobs	280
r2	0.782927
logloss	0.453875
mean_per_class_error	0.336996

Combine predictions with frame

Fig. 5. Distributed Random Forest Model with 75% of the dataset



PREDICTION - CM - CONFUSION MATRIX

I am actively looking for a job	*I am not interested in new job opportunities*	*I'm not actively looking, but I am open to new opportunities*	*Error*	*Rate*
0	0	32	1.0	32 / 32
0	75	0	0	0 / 75
0	1	159	0.0063	1 / 160
0	76	191	0.1236	33 / 267

Fig. 6. Confusion Matrix for prediction on the test data

The accuracy of the model with the test data can be computed as: ((75 + 159)) / ((32 + 75 + 160)) x 100 = 87.64%. With the results demonstrated in this paper, it can be inferenced that most developers may be looking towards new job prospects as Data Scientists based on the trends in the programming languages and databases which are widely used in the field of Big Data Analytics. However, the prediction model using the different parameters show that the developers may not be looking to move to new job environments and would prefer staying in their current company or organisation. This would be an indication that developers are most probably looking forward to bring added value to their current companies/organisations as Big Data Analytics would start to be adopted.

IV. Conclusion and Future Works

The work in this paper, analyses the data from the survey conducted with the numerous developers on Stack Overflow in order to gain insights on the directions of programming languages, databases, and the job seeking status of the developers. Results show that the trend is developers want to use more the programming languages and databases used on cloud platforms for Big Data Analytics. Additionally, a Distributed Random Forest model with 87.64% accuracy, for predicting the job seeking status of developers shows that the developers may not be looking to move to new job environments and would prefer staying in their current company or organisation. This would be an indication that developers are most probably looking forward to bring added value to their current companies/organisations as Big Data Analytics would start to be adopted.

References

[1] M. Chambers, C. Doig and I. Stokes-Rees, Breaking Data Science Open (How Open Data Science is Eating the World), CA: O'Reilly Media Inc., 2017.

[2] L. Columbus, “IBM Predicts Demand For Data Scientists Will Soar 28% By 2020,” IBM, 13 May 2017. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#7394fa4b7e3b. [Accessed 26 December 2017].

[3] V. Granville, “Data Science Central,” 14 December 2016. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/data-science-machine-learning-iot-2017-predictions. [Accessed 26 December 2017].

[4] Environmental Science, “What Is a Data Scientist?,” 2017. [Online]. Available: https://www.environmentalscience.org/career/data-scientist. [Accessed 26 December 2017].

[5] A. Rosenblum, “The Tools of Big Data Science: The Technologies & Languages of Statistical Analysis,” Business-2-Community, 19 March 2016. [Online]. Available: https://www.business2community.com/big-data/tools-big-data-science-technologies-languages-statistical-analysis-01483461. [Accessed 26 December 2017].

[6] Gartner Inc., “Gartner Says More Than 40 Percent of Data Science Tasks Will Be Automated by 2020,” Gartner Inc., Sydney, Australia, 2017.

[7] G. Piatetsky, “Forrester vs Gartner on Data Science Platforms and Machine Learning Solutions,” KDnuggets, April 2017. [Online]. Available: https://www.kdnuggets.com/2017/04/forrester-gartner-data-science-platforms-machine-learning.html. [Accessed 26 December 2017].

[8] M. S. Farooq, S. A. Khan, F. Ahmad, S. Islam and A. Abid, “An Evaluation Framework and Comparative Analysis of the Widely Used First Programming Languages,” Plos One, vol. 9, no. 2, pp. 1-25, 2014.

[9] B. Vassilev, R. Louhimo, E. Ikonen and S. Hautaniemi, “Language-Agnostic Reproducible Data Analysis Using Literate Programming,” Plos One, vol. 11, no. 10, pp. 1-14, 2016.

[10] Kaggle, “Kaggle,” 2017. [Online]. Available: https://www.kaggle.com/stackoverflow/so-survey-2017/data. [Accessed 27 December 2017].

[11] KNIME AG, “H2O Random Forest Learner,” H2O, [Online]. Available: https://yifydownloads.com/iron-sky-2012/. [Accessed 19 May 2018].

[12] E. S. Walsh, B. J. Kreakie, M. G. Cantwell and D. Nacci, “A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system,” PloS ONE, vol. 12, no. 7, pp. 1-18, 2017.

[13] I. Pointer, “Which freaking big data programming language should I use?,” Info World, April 2016. [Online]. Available: https://www.infoworld.com/article/3049672/application-development/which-freaking-big-data-programming-language-should-i-use.html. [Accessed May 2018].