Paper 9700-2016

Assessing Undergraduate and Postgraduate Hard and Soft Skills in Analytics and Data Science Courses

Richard J Self, University of Derby, UK

r.j.self@derby.ac.uk

ABSTRACT

Traditional approaches to assessing undergraduate assignments in the field of software related courses, including Analytics and Data Science courses, involve the course tutors in reading the students' code and getting the students to physically demonstrate their artefacts. However, this approach tends only to assess the technical skills of solving the set task. It generally fails to assess the many soft skills that industry is looking for, as identified in the e-skills UK (Tech Partnership) / SAS®[1] report from Nov 2014 and the associated infographic poster.

This paper will describe and evaluate the effectiveness of a different approach to defining the assessment task and formatively and summatively assessing the work of the students in order to effectively evaluate and mark both the soft skills, including creativity, curiosity, storytelling, problem solving and communication, as well as the technical skills. This approach works effectively at all levels of undergraduate and masters courses.

INTRODUCTION

Employers want re-assurance that graduates from universities in the fields of Analytics and Data Science have both technical and soft skills.

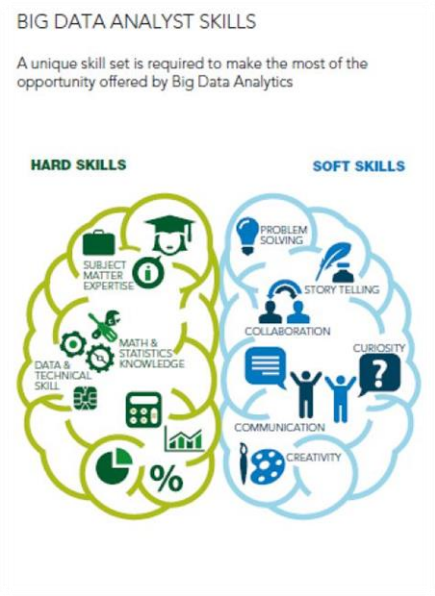

Figure 1 – Big Data Analyst Skills

- [1] SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
- Other brand and product names are trademarks of their respective companies.

Many assessments concentrate on the technical skills, shown in the left brain image in Figure 1 above, which was presented in the Tech Partners / SAS® infographic dated Nov 2014. However, employers are demanding that Universities also develop students' soft skills, shown in the right brain in Figure 1 above.

The University of Derby has developed novel approaches to both defining the task and the assessment rubrics which provide strong emphasis on both sets of skills. As examples, early modules in the students' careers at the University of Derby are assessed by a portfolio of tasks which use Computer based Tests to assess technical SAS® skills and research and presentation tasks to evaluate their soft skills.

Courses that occur later in the programs concentrate the assessment focus towards the soft skills and rely on the students to reflect on some of the relevant technical skills that were involved in the task through the medium of a clearly specified structure for a presentation that critically evaluates aspects such as:-

- Requirements specification and justification of the project (20% weighting)
- Design, development and implementation (20% weighting)
- Insights gained from the project (40% weighting)
    - Technical integration issues
    - Data sourcing / cleansing / integration issues
    - Analytical insights
    - Data for Humanity principles embodied

In order to create the greatest engagement by the students in the courses, the overall assessment tasks are specified in quite general terms and require each student to identify a specific set of data which is of interest to them, based on directed research.

The techniques presented in Paper 3203-2015, Learning Analytics to evaluate and confirm Pedagogic Choices, at SAS GF 2015, have been used to evaluate the impact of this approach.

THE COURSES

Three modules have been used to pilot the approach at first year, final year and MSc levels.

The first is a first year second, semester module called "I**ntroduction to Data Analysis**" which gets the students to learn SAS, using the materials from the Base SAS course as self-study.

The second module is a final year, first semester module called "**Emerging IT Product Development**" introduces the students to the IBM Bluemix and Watson Analytics product set, via a couple of IBM presented tutorials, on-line videos and self-directed exploration.

The third is a Master's level module called "**Business Analytics with SAS**" which gets the students to learn SAS, using the materials from the Base SAS course as self-study.

INTRODUCTION TO DATA ANALYSIS MODULE SPECIFICATION

Module Content

The module will explore the following topics by learning one or more industry-standard tools for data analysis:

- Basic descriptive statistics and their meaning
- Data analysis tools
- Data formats and importing data
- Validating and cleaning data
- Manipulating data and combining data sets
- Generating reports
- Generating graphs and graphics

Module Learning Outcomes

On successful completion of the module, students will be able to

1. Understand and apply basic statistical and data analysis techniques.
2. Effectively design and implement simple systems for performing basic data analysis and reporting.
3. Use data analysis tools to communicate results effectively to others.

EMERGING IT PRODUCT DEVELOPMENT MODULE SPECIFICATION

Module Content

Emerging technologies have become an area of growing importance both in research and industry. This is mainly due to the Future Internet as a means to foster innovation towards creating and sustaining competitive advantage. These emerging IT products often need to been seen both as distinct services but also as integrated services functioning in a customised, synergetic and ad-hoc manner. The module will expose students to these cutting edge IT developments including but not limited to web services, data mashups, crowd sourcing, social networking, grid and cloud computing, context-aware, Internet of Things and other next generation technologies as they emerge and as the tools to support current challenges. Apart from the developmental aspects, students will be prompted to identify opportunities and provide future direction for emerging IT utilisation. Within this context, a particular focus is on developing the critical thinking to evaluate aforementioned developments and on applying rational and appropriate methods to selecting these when produced with a variety of choices. While, these technologies will be analysed to determine their applicability and appropriateness for and impact from their adoption in the real world, the Sustainable Information and Corporate Governance module will take a complementary role by critically evaluating their sustainable governance in terms of ethics, implications and professionalism.

Module Learning Outcomes

On successful completion of the module, students will be able to:

1. Be conversant and demonstrate higher analytical skills and critical understanding about developments in emerging technologies, and demonstrate the ability to make objective, rational decisions towards their adoption in the real world
2. Be able to identify and apply emerging technologies tools towards the development of an advanced IT product as well as justify and evaluate their developmental decisions

BUSINESS ANALYTICS WITH SAS MODULE SPECIFICATION

Module Content

The module will consider topics such as:-

- The principles of data analysis and normalization as the foundation of the “store once and once only” principle for data integrity
- The requirements for optimization of data management for the conflicting needs of real-time and MIS / EIS and BI purposes
- The impact of User-centered analysis and participatory development on the development data structures and stores and systems that meet the needs of the organization
- The role, application and benefits of prototyping in the rapid development of systems, using prototyping, RAD and Agile approaches
- Techniques and practices for effective data analysis and presentation

Together with a range of current, leading edge topics and technologies that affect data management and usage.

Module Learning Outcomes

On successful completion of the module, students will be able to:

1. Identify and critically evaluate the key requirements for effective data management, analysis and presentation to meet users' needs
2. Develop a small management information reporting system that meets the users' needs using a prototyping / Agile approach in the SAS programming environment

TEACHING, LEARNING AND ASSESSMENT STRATEGIES

Each of these three courses introduce the students to new programming languages, require them to both learn the system and then create some sort of artefact in that language and then to critically review their project and their achievements.

TEACHING AND LEARNING STRATEGY

The pedagogic foundation that underpins this approach is a change from the comparatively traditional perspective of the "Academic as Domain Expert" to an approach of the "Academic as source of Learning to Learn Skills". A consequence of this changed perspective is that the, so-called, teaching approach is no longer to concentrate on delivering the domain knowledge, facts and answers but on clarifying domain concepts, relationships and sets of important questions that the learners must be aware of and understand. This provides great freedom to use as much of the allocated contact time as possible or necessary for one-to-one and very small group formative discussion and feedback; there is little or no need to dedicate much time to delivering the domain facts which can be easily researched and read by the students, provided that suitable signposting is provided. It is an approach that stops trying to fill leaky buckets with information and replaces it with lighting the fires of inspiration and enquiry (to amend Plutarch (100 A.D.)).

It also means that formal teaching of programming languages and systems is not necessary or productive. It is certainly not engaging to the students. Instead, the approach takes the stance that learning the syntax and grammar should be a self-directed task, leaving the contact time for solving problems and difficulties with the learning materials made available to the students, where available, or they should be directed to search for expert learning materials and code libraries that are publicly available via the internet. This turns out to replicate what our graduates will be doing once the enter employment, where the probability of being able to afford to attend formal courses will be very limited in many SME organizations where most of our graduates gain employment.

As a consequence, for the Base SAS related courses, we use the standard "Base SAS" 20 chapters' materials as a self-study program which they complete during workshops over the first 5 weeks of the semester. Their grasp of the technicalities is then confirmed via 4 Computer Based Tests which contribute to their overall score for the portfolio.

For the IBM Bluemix and Watson Analytics topics for the third year module, we arranged for support at two seminar workshops from an IBM expert which was then followed up with a workshop towards the end of the semester to solve assessment artefact issues. Following these three sessions, the students explored the systems and used online help to develop their understanding and skills.

ASSESSMENT STRATEGY

It has become apparent that student engagement can be significantly improved if the students are allowed the freedom to negotiate an assessment topic that is close to their personal interests.

Sources of Data

As a result, each of the undergraduate modules provides a broad topic and challenge within which the students must identify data that is of personal interest and significance and meets certain general criteria in terms of size and complexity. The first year challenge requires two data sets containing at least 2000 rows each that have some relationship. The third year challenge is to find larger sets of data, approaching "biggish" data. One student used a dataset of 49M rows based on the UK National Crime statistics files.

The Master's module assessment task is based on furthering the academic's own personal academic research, so they are directed to a specific data set from the Microsoft GeoLife[2] project, containing approximately 1.5GB and some 17000 GPS tracked journeys, mainly in Beijing, containing approximately 23M data points.

Technical Analytics Challenge

Each of the modules presents the analysis challenge to the students in a different way, as can be seen below as part of an overall portfolio of tasks which might include outline requirements specifications, design criteria justifications, critical evaluation of the project execution, etc.

Introduction to Data Analysis (first year)

As part of the summative assessment for this module, you are required to apply the skills and knowledge that you have acquired in the taught classes, to build a demonstration data analysis system.

Assessment of the Task

You will evidence your work by creating a short video presentation of **no more than 4 minutes duration**,that demonstrates a walk-through of your system.

The **mandatory** requirements of the system are as follows:

R1. The system will analyse data from at least two separate sources.

R2. Each data source must contain at least 2000 records,

R3. You will use Base SAS to create datasets for subsequent analysis.

R4. The system will have the functionality to enable static reporting.

R5. The system will have discrete stages for extract, transform and load.

Optional functionality:

O1. The reporting may include graphical visualisation.

O2. An interactive dashboard that permits query-on-demand reporting.

Emerging IT Product Developments (final year)

- You will need to select a suitable source (or sources) of Open Data that you will input to one or more Bluemix products.
- You will identify an interesting problem that can be found within the data and which suitable analytics will enable you to gain valuable insights into the problem.
- You will then develop a suitable requirements specification, which will also justify the choice of data and the problem that you want to gain insights about.
- You may need to carry out some data cleansing as you load the data.
- You will then use the functionality of the Bluemix products that you have chosen in order to perform some relevant analytics from which you can develop valuable insights.

Assessment of the Task

You will create a presentation, using PowerPoint with an added voice-over (using the Slideshow / Record Slide Show capability), which will last for **15 minutes +/- 1 minute**. If your presentation is less than 14 minutes of over 16 minutes then you will lose 50% of the available presentation marks, that is 10 marks overall of the assessment.

[2] GeoLife is to be found at http://research.microsoft.com/en-us/projects/geolife/default.aspx

This presentation be a critical reflection of the project and will cover

- Requirements specification and justification of the project (20% weighting)
- Design, development and implementation (20% weighting)
- Insights gained from the project (40% weighting)
    - Technical integration issues
    - Data sourcing / cleansing / integration issues
    - Analytical insights
    - Data for Humanity principles embodied

BI With SAS (MSc)

Overall, you will specify, design and build a small management information analysis and presentation system to analyse some of the Journey data that can be found on the S:drive at S:\Common Area\SAS\Location Data\ in the Geolife Trajectories Zip File. The .plt files are actually CSV files which can initially be read in Excel. You will need to omit the first 6 rows of data which have no relevance and import some into SAS.

Assessment of the Task

The second part of the portfolio which will be assessed in my office (E516) on Tuesday 12 April 2016 in class with a short demonstration of the system.

The critical evaluation of the overall project will be a short 1000 to 1500 word report in the Springer LNCS format to be submitted to Turnitin on Monday 25 April by 11:59pm and will be marked in my office on Tuesday 26 April 2016 to a schedule to be specified.

ASSESSMENT CRITERIA

These are designed to be related to the standard university criteria for the relevant level and also should be both understandable by students (in terms of related to their own experience) and also robust, by strong connection to relatively objective external criteria, so that high grade profiles cannot be re-normalized to a lower grade average.

The robustness is obtained by relating the highest grade bands to external publishing and conference standards. Most academics are involved in publications and conference papers and presentations. As a result, it is generally easy to agree with both internal and external moderation that the high standars are appropriate and fair.

Examples are given below of the top level criteria for the soft skill assessment of the technical challenge artefacts.

Introduction to Data Analysis (first year)

This just uses three criteria to assess the competence of the solution of the technical challenge, as presented in the video

Criterion	95	85
Design and implement a working data analysis application	Outstanding design with innovative features, some attempt to optimise code	Excellent example of a prototype system, novel architecture

Utilise data analysis techniques for knowledge discovery	Lateral thinking demonstrated with respect to data analysis techniques selected	Comprehensive analysis that is automated
Specify and implement a reporting mechanism	Outstanding visualisation that is dynamic and fully automated/interactive	Excellent visualisation that may be automated in part

Emerging IT Product Developments (final year)

Presentation	Criteria	Requirements Specification and Justification of Project	Design, Development & Implementation	Insights Technical Data Analytical Principles
(20%)	Weighting	(20%)	(20%)	(40%)
Novel topic that has been clearly articulated and also grounded in the existing literature. Presentable in an international workshop and demonstrates **SFIA level 6/7 capabilities (Strategy / Inspire)**	1st (95%)	Exceptionally lucid argument, with logical and novel conclusions, thorough critical evaluation	Exceptionally lucid argument, with logical and novel conclusions, thorough critical evaluation	Exceptionally lucid argument, with logical and novel conclusions, thorough critical evaluation

BI With SAS (MSc)

The artefact is assessed by two criteria that are, effectively, the only criteria for defining the challenge and are used by the students to calibrate what level of effort they will invest in interesting and challenging solutions to the system.

Percentage Band	**Coding and use of SAS – 50%**	**Data – 50%**
95%	Relevant SAS Source code pre-written and stored in SAS system. Code modules chosen and run from a menu system. Data selection menu driven with complex options. Graphic data presentation, good choice of output styles which are menu driven.	Uses pre-imported data in a saved master file using at least 1000 journeys data. Optimises SAS storage structures to the types of questions provided. Optimises use of permanent and work file structures.
85%	Relevant SAS Source code pre-written and stored in SAS system. Code modules chosen and run	Uses pre-imported data in a saved master file using at least 500 journeys data.

	from a menu system. Data selection menu driven with complex options. Graphic data presentation, good choice of output styles.	Optimises SAS storage structures to the types of questions provided.
75%	Relevant SAS Source code pre-written and stored in SAS system. Code modules chosen and run from a menu system. Data selection code pre-written with complex options. Graphic data presentation, good choice of output styles.	Uses pre-imported data in a saved master file using at least 300 journeys data.

GENERAL COMMENT

It should be noted that the students will need to undertake considerable research in order to find the relevant code libraries and hints and tips, in order to achieve the higher grade levels.

As an example, for the MSc artefact, they will need to find out how to import large numbers of files from many folders.

OVERALL RESULTS

The majority of all students achieve high grades. According to the student feedback, this is partly due to the freedom to find their own data or their own analytical techniques and tools. They say that this allows them to connect the task and data to their own personal interests and causes them to invest much more effort in the research and development of the technical challenge artefact.

An additional result of the continuous use of soft skill based assessment is that by the final year, the BSc IT students are far more capable of communicating effectively, identifying the key insights and creating and presenting a clear story. The final year students state that they feel far more confident in their ability to create a coherent story and then tell that story to the relevant audience, whether as an assessment or in a job interview or in the workplace.