AMEO 2015 - A dataset comprising AMCAT test scores, biodata details and employment outcomes of job seekers

Varun Aggarwal
Aspiring Minds
varun@aspiringminds.com

Shashank Srikant
Aspiring Minds
shashank.srikant@aspiringminds.com

Harsh Nisar
Aspiring Minds
harsh.nisar@aspiringminds.com

ABSTRACT

More than a million engineers enter the global workforce every year. A relevant question is what determines the jobs and salaries these engineers are offered right after graduation. Previous studies have shown the influence of various factors such as college reputation, grades, the field one specializes in and market conditions for specific industries. An important input which such analyses do not have is a standardized measures of job skills done at the time of completion of studies. We present here Aspiring Minds' Employability Outcomes 2015 (AMEO 2015), a unique dataset which provides engineering graduates' employment outcomes (salaries, job titles and job locations) together with standardized assessment scores in three fundamental areas - cognitive skills, technical skills and personality. Coupled with biodata information, AMEO 2015 provides an opportunity for a unique and comprehensive study of the entry level labor market. The data could be used to make an accurate salary predictor, but also understand what influences salary and job titles in the labor market. In this paper we describe the details of the dataset and discuss a spectrum of questions around meritocracy in labor markets, biases in labor selection and other prevalent market forces it can help uncover and answer. You can download the dataset at: http://research.aspiringminds.com/resources/

CCS Concepts

•**Information systems** → *Information systems applications;* •**Computing methodologies** → **Machine learning;** •**Applied computing** → *Economics;*

Keywords

dataset; labour market; data mining

1. INTRODUCTION

According to OECD calculations, there will be more than 200 million 25-34 year-olds with higher education degrees across all OECD and G20 countries by the year 2020 [7]. In spite of such numbers, the current education to employment ecosystem is hardly transparent and offers little insights in labor migration, compensation and job availability to help equip both, the job seeker and job provider with the right information. Answers to questions like what really predicts a college graduate's salary and what distribution of skills exist in the entry-level workforce across various geographies can immensely help both actors in the ecosystem. However, in order to answer such questions, one needs to find what links each graduate's biodata information to her/his employment outcome - the job title offered, location of the job and the salary offered with his/her skills. This would facilitate a more systematic study on how the workforce is allotted jobs in current labor markets. In the process, it would also identify the existence of biases in the current labor markets, such as salary disparity between specific groups like gender and those graduating from institutions with higher reputation. Unfortunately, no such data set exists which consolidates scores on a post-education job skills test with biodata information and eventual employment outcomes. To address this lack of information, we present Aspiring Minds' Employment Outcomes 2015, a unique dataset of engineers entering the workforce. At Aspiring Minds, a global job skills credentialing company, we have the unique advantage of assessing over a million undergraduates and post graduates in three fundamental areas - cognitive skills, technical skills and personality. For a subset of those assessed on our platform, we tracked details of which companies they were hired by, what salaries they were offered, what job titles they were provided and which locations they were assigned to. Such a dataset promises to provide a holistic picture of the dynamics of entry-level labor markets.

There exist other similar publicly available datasets containing demographic information and employment outcomes of a workforce. [8][5] However, they have the limitation of not connecting the entire picture. These datasets either miss post-education standardized assessments scores, employment outcomes or both. AMEO contains both of these along with demographic information, making it a dataset useful for substantive studies.

CODS '16, March 13 - 16, 2016, Pune, India

ISBN 978-1-4503-4217-9/16/03…$15.00

DOI: http://dx.doi.org/10.1145/2888451.2892037

2. DATASET DESCRIPTION

For every engineer, it provides anonymized biodata information along with their respective skill scores and employment outcome information. For detailed information on each attribute refer to Table 2 in Appendix A.

Specifically, the following information is available for every

Table 1: Basic description of dataset

Data Set Characteristics	Multivariate
Number of Instances	3998
Total Attributes	38
Attribute Characteristics	Date-time, Ordinal & Integer
Missing Values?	Yes

engineer:

1. Scores on Aspiring Minds' AMCAT [1] - a standardized test of job skills. The test includes cognitive, domain and personality assessments.
2. Personal information like gender and date of birth.
3. Pre-university information like 10^{th} and 12^{th} grade marks, board of education and 12^{th} grade graduation year.
4. University information like GPA, college major, college reputation proxy, graduation year and college location.
5. The following employment outcome information is available for every engineer:
   - First job annual salary
   - First job title
   - First job location
   - Date of joining and leaving of first job

3. DATASET COLLECTION

Over a million students, on completing university education, take AMCAT every year to get their skills credentialed. We reached out to nearly 80,000 such students who had graduated in 2015 to survey them for their employment outcomes. We focused our reach out only on engineers since they were the subject of our study. All those reached out to were explained of the purpose of collecting this data. Each respondent filled up a form which asked the number of jobs s/he has had since leaving university. Job details against each such job was collected. For job titles, we used a curated list of 1200 commonly found job titles on the internet and we asked respondents to either pick one from the list or fill a custom job title. We had a similar list of cities and states for the respondents to pick. The dataset was finally collated on details from only the first jobs of the respondents.

4. USAGE OF DATASET

AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the *National Employability Report*, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset [3]. It was also analyzed as part of two data challenges - at ASSESS 2015 [6], a workshop on data-driven assessments held at ICDM 2015 and at IKDD CODS 2016 [2]. As part of the challenge, researchers from industry and academia predicted annual salaries of engineering graduates. The challenge also had them interpret the factors determining salaries and had them visualize the dataset to infer insights. Further, AMEO is being used in an online course to teach hands on machine learning and visualization techniques [4]. The dataset has a variety of input formats, including semi-structured text in the form of job titles, making it a rich instructional tool which students and professionals can readily relate to and learn from. It can also be used to teach quantitative social science methods.

5. CONCLUSION

We present in this work details of Aspiring Minds' Employability Outcomes 2015 (AMEO 2015), the first dataset to provide engineers' employment outcomes (salaries, job titles and job locations) together with standardized assessment scores in three fundamental areas - cognitive skills, technical skills and personality. The dataset also contains anonymized biodata, making it a rich source for a comprehensive study of the entry-level labor markets. We see it being used in analyses which can inform policies on personnel selection and higher education. It can also be used to model a variety of data-driven systems which can efficiently match job seekers and job providers. Systems like these include accurate salary predictors, automated job counselors and visualizations on how skills are distributed across geographies. We look forward to this dataset being utilized in the design of a truly merit-driven labor market.

6. REFERENCES

[1] Aspiring minds. http://www.aspiringminds.com.

[2] ACM IKDD CODS. Data challenge, March 2016. http://ikdd.acm.org/Site/CoDS2016/.

[3] Aspiring Minds. National employablity report - engineers annual report, 2015. http://www.aspiringminds.com/research-reports.

[4] Aspiring Minds. Introduction to machine learning, 2016. https://lms.aspiringminds.in/home.

[5] Government of India, Ministry of Statistics and Programme Implementation. Employment and Unemployment : NSS 61st. round, 2004. [Online; accessed 15-February-2016].

[6] ICDM. ASSESS: Data Mining for Educational Assessment and Feedback, 2015.

[7] OECD. Education indicators in focus. pages 2–3, May 2012.

[8] United States Department of Labor, Bureau of Labor Statistics (BLS). National compensation survey, 2015. [Online; accessed 15-February-2016].

APPENDIX

A. DESCRIPTION OF ATTRIBUTES

Table 2 on page 3 contains detailed description of each attribute in the dataset.

Table 2: Description of Attributes

Input	Description	Comments
ID	A unique ID to identify a candidate	
Salary	Annual CTC offered to the candidate (in INR)	Self reported by the candidate.
DOJ	Date of joining the company	Self reported by the candidate.
DOL	Date of leaving the company	A value of "present" means the candidate continues to work at the company at the time of collecting this information
Designation	Designation offered in the job	Textual/non-standardized. Self reported by the candidate.
JobCity	Location of the job (city)	Textual/non-standardized. 262 Unique values. -1 represents missing data.
Gender	Candidate's gender	m denotes Males and f denotes Females
DOB	Date of birth of candidate	
10percentage	Overall marks obtained in grade 10 examinations	Domain of values: [0,100]
10board	The school board whose curriculum the candidate followed in grade 10	India has several boards of education which follow their own course curricula. Schools are affiliated to one of these boards.
12graduation	Year of graduation - senior year high school	
12percentage	Overall marks obtained in grade 12 examinations	Domain of values: [0,100]
12board	The school board whose curriculum the candidate followed in grade 12	
CollegeID	Unique ID identifying the college which the candidate attended	Multiple candidates may belong to the same university/college in this dataset. College names have not been disclosed to maintain privacy.
CollegeTier	Tier of college	Categorical. Domain of values: {1,2}. Computed from the average AMCAT scores obtained by the students in the college. Colleges with an average score above a threshold are tagged as 1 and others as 2.
Degree	Degree obtained/pursued by the candidate	Categorical. Clean and standardized.
Specialization	Specialization pursued by the candidate	Categorical. Clean and standardized. A total of 46 specializations are represented.
CollegeGPA	Aggregate GPA at graduation	This is the raw information submitted by candidates. Some have submitted percentages while others have posted on a 10-point scale.
CollegeCityID	A unique ID to identify the city in which the college is located in.	
CollegeCityTier	The tier of the city in which the college is located	Categorical. Domain of values: {1,0}. 1 represents a tier-one city while 0 represents a tier-2 city. Tier is decided based on population.
CollegeState	Name of the state in which the college is located	Categorical. A total of 26 states are represented. No missing values. The provided data set does not accurately capture state-wise distributions.
GraduationYear	Year of graduation (Bachelor's degree)	
English	Scores in AMCAT English section	Domain of values: [100,900]
Logical	Score in AMCAT Logical ability section	
Quant	Score in AMCAT's Quantitative ability section	
Domain	Scores in AMCAT's domain module	Since different candidates give different domain-specific tests, we report here the percentile of the candidates in their respective tests. Domain of values: [0,1]. This is an optional section for the candidates. Those opting out of it get a score of -1.
ComputerProgramming	Score in AMCAT's Computer programming section	Domain of values:[100,900] This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. If the candidate's domain-specific test is the same as one of these tests, domain percentile will be based on the score presented here.
ElectronicsAndSemicon	Score in AMCAT's Electronics & Semiconductor Engineering section	
ComputerScience	Score in AMCAT's Computer Science section	
MechanicalEngg	Score in AMCAT's Mechanical Engineering section	
ElectricalEngg	Score in AMCAT's Electrical Engineering section	
TelecomEngg	Score in AMCAT's Telecommunication Engineering section	
CivilEngg	Score in AMCAT's Civil Engineering section	
conscientiousness	Scores in one of the sections of AMCAT's personality test	The scores are sampled from a distribution with mean 0 and standard deviation 1 and represent the Big 5 personality traits.
agreeableness	Scores in one of the sections of AMCAT's personality test	
extraversion	Scores in one of the sections of AMCAT's personality test	
nueroticism	Scores in one of the sections of AMCAT's personality test	
openess_to_experience	Scores in one of the sections of AMCAT's personality test	