Probability of an Occupation Being Analytical

Author: Nazli Rafei Dehkordi
Email: rafeidehkordi.n@northeastern.edu

1. Labeling

To hand label the response variable, I randomly selected 150 variables and assign 1 to those decided as Analytical and 0 to any job decided as not analytical. label assignments were based on eyeballing the O∗NET tasks and job description of each occupation[1].

The hand-labelling of the occupations was made by going through their job description and see if job is done through any analytical skills. Skills like Idea Generation and Reasoning Abilities ,Critical thinking, Data and information analysis, Research , communication, Problem-solving, creativity considered as analytical skills based on indeed website [2] and my own judgment.
Same criteria were used for objective assignment of being analytical or not.

Thus, I assumed a task Analytical if the job cannot be done without having analytical skills. So "level" from scale table was chose to define as measure of how much of these skills we have for the corresponding job. As reported in Table 1, 22 attributes were identified as describing these skills. These attributes as it is shown in Table 1, are defined in Content Model Reference table and they have 5 levels defining them.

Table 1. Content Model Reference data example

Element ID	Element Name	Description
1	Worker Characteristics	Worker Characteristics
1.A	Abilities	Enduring attributes of the individual that influence performance
1.A.1	Cognitive Abilities	Abilities that influence the acquisition and application of knowledge in problem solving
1.A.1.a	Verbal Abilities	Abilities that influence the acquisition and application of verbal information in problem solving
1.A.1.a.1	Oral Comprehension	The ability to listen to and understand information and ideas presented through spoken words and sentences.

In some tables like skills, we have element id related to these skills defined as the higher level like (1.A.1.a.1) but in abilities element ids are defined with four characters like (1.A.1.a). 5-character element ids encompass features of 4 characters. So, I used first 4 character of these in attributes in all tables which are rows in gray. Selected list of attributes consists of Idea Generation and Reasoning Abilities, Innovation, Analytical Thinking, Active Learning, Complex Problem Solving, Information and Data Processing, Reasoning and Decision Making, total of 8 attributes (Table 3).

Table 3. O∗NET element that serve as indicators of a job being analytical.

Element ID	Element Name	Description
1.A.1.b	Idea Generation and Reasoning Abilities	Abilities that influence the application and manipulation of information in problem solving
1.A.1.b.1	Fluency of Ideas	The ability to come up with a number of ideas about a topic (the number of ideas is important, not their quality, correctness, or creativity).
1.A.1.b.2	Originality	The ability to come up with unusual or clever ideas about a given topic or situation, or to develop creative ways to solve a problem.
1.A.1.b.3	Problem Sensitivity	The ability to tell when something is wrong or is likely to go wrong. It does not involve solving the problem, only recognizing that there is a problem.
1.A.1.b.4	Deductive Reasoning	The ability to apply general rules to specific problems to produce answers that make sense.
1.A.1.b.5	Inductive Reasoning	The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events).
1.A.1.b.6	Information Ordering	The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations).
1.A.1.b.7	Category Flexibility	The ability to generate or use different sets of rules for combining or grouping things in different ways.
1.C.7.a	Innovation	Job requires creativity and alternative thinking to develop new ideas for and answers to work-related problems.
1.C.7.b	Analytical Thinking	Job requires analyzing information and using logic to address work-related issues and problems.
2.A.2.a	Critical Thinking	Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems.
2.A.2.b	Active Learning	Understanding the implications of new information for both current and future problem-solving and decision-making.
2.B.2.i	Complex Problem Solving	Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions.
4.A.2	Mental Processes	What processing, planning, problem-solving, decision-making, and innovating activities are performed with job-relevant information?
4.A.2.a	Information and Data Processing	How is information processed to perform this job?
4.A.2.a.1	Judging the Qualities of Objects, Services, or People	Assessing the value, importance, or quality of things or people.
4.A.2.a.2	Processing Information	Compiling, coding, categorizing, calculating, tabulating, auditing, or verifying information or data.

4.A.2.a.3	Evaluating Information to Determine Compliance with Standards	Using relevant information and individual judgment to determine whether events or processes comply with laws, regulations, or standards.
4.A.2.a.4	Analyzing Data or Information	Identifying the underlying principles, reasons, or facts of information by breaking down information or data into separate parts.
4.A.2.b	Reasoning and Decision Making	What decisions are made and problems solved in performing this job?
4.A.2.b.1	Making Decisions and Solving Problems	Analyzing information and evaluating results to choose the best solution and solve problems.
4.A.2.b.2	Thinking Creatively	Developing, designing, or creating new applications, ideas, relationships, systems, or products, including artistic contributions.

2. Modeling

After hand labeling 150 randomly selected samples from our data, 46 of jobs from this sample are labeled as analytical and one has been assigned to them, and 104 labeled as non-analytical and 0 has been assigned to them. As for modeling, we are dealing with an imbalance data set , and we have a binary classification.

to estimate the probability of an occupation being analytical, logistic regression with and without considering weight, Gaussian Process Classifiers (GPC) with evaluation different kernels both for imbalanced binary classification are tested. and for. Both these techniques are specifically used to get the probability of an event occurring, not just the predicted classification.

Rando forest was a candidate to choose however, probability of the classes in random forest is more an score rather than "true" probability, and it needs further calibration of the method to make sure it is the probability of interest [3]. On the other hand Logistic regression and Gaussian processes are able to predict highly calibrated probabilities [4–6].

3. Modeling

As we have imbalanced data set accuracy is not a good measure of performance, and evaluation metrics like ROC-AUC curve are a good indicator of classifier performance. The Higher the ROC-AUC score, the better the model is at predicting 0s as 0s and 1s as 1s.

Table 4. Performance of various classifiers; best performances in bold.

Model	Accuracy Score on Test set	Area Under Curve	Recall score
Logistic Regression	0.755	0.700	0.533
Weighted Logistic Regression	0.777	0.800	0.866
Gaussian Process Classifier	**0.785**	**0.812**	**1.00**

Result for performances of our models is tabulated in Table 4, all measuring scores are better in Gaussian Process Classifier, model. So, I continued prediction with GPC method. GPC model did a decent job on prediction jobs in whole data set for both classes of being analytical and not analytical jobs.

Figure 1 shows probability distribution of model over whole dataset, as we see distribution of probabilities are not highly concentrated around border threshold, which is 0.5 in this case, and reflecting the fact that model is strong in prediction.

Figure 1. Histogram of predicted probabilities

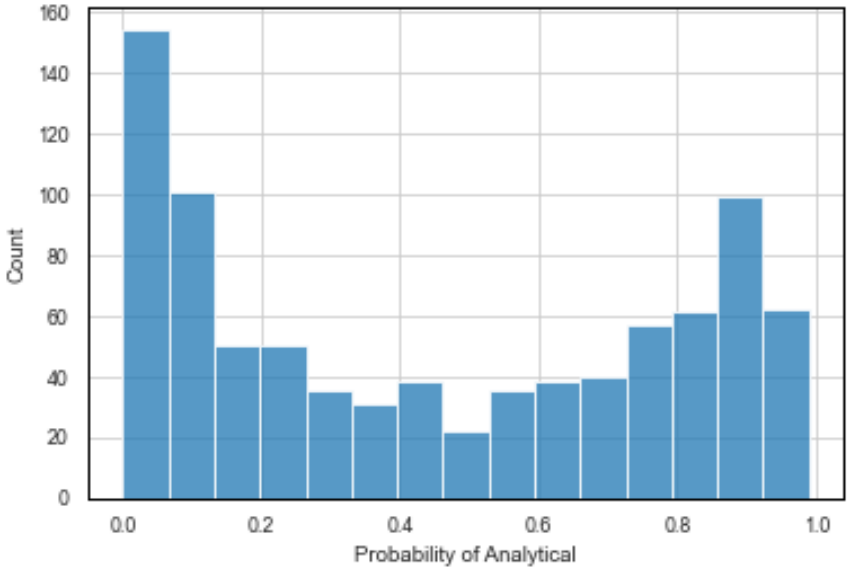

Table 5 shows occupations predicted as analytical that was not expected to be predicted as analytical, and Table 6 represents occupations that expected to be analytical and were not predicted as analytical.

Table 5. Occupations come up as Analytical which were not expected to be Analytical

Job	Probability	Predicted	Description
Buyers and Purchasing Agents, Farm Products	0.56	1	Purchase farm products either for further processing or resale. Includes tree farm contractors, grain brokers and market operators, grain buyers, and tobacco buyers. May negotiate contracts.
Purchasing Agents, Except Wholesale, Retail, and Farm Products	0.76	1	Purchase machinery, equipment, tools, parts, supplies, or services necessary for the operation of an establishment. Purchase raw or semifinished materials for manufacturing. May negotiate contracts.
Customs Brokers	0.53	1	Prepare customs documentation and ensure that shipments meet all applicable laws to facilitate the import and export of goods. Determine and track duties and taxes payable and process payments on behalf of client. Sign

			documents under a power of attorney. Represent clients in meetings with customs officials and apply for duty refunds and tariff reclassifications. Coordinate transportation and storage of imported goods.
Human Resources Specialists	0.61	1	Recruit, screen, interview, or place individuals within an organization. May perform other activities in multiple human resources areas.
Credit Counselors	0.67	1	Advise and educate individuals or organizations on acquiring and managing debt. May provide guidance in determining the best type of loan and explain loan requirements or restrictions. May help develop debt management plans or student financial aid packages. May advise on credit issues, or provide budget, mortgage, bankruptcy, or student financial aid counseling.
Health Specialties Teachers, Postsecondary	0.87	1	Teach courses in health specialties, in fields such as dentistry, laboratory technology, medicine, pharmacy, public health, therapy, and veterinary medicine.
Nursing Instructors and Teachers, Postsecondary	0.88	1	Demonstrate and teach patient care in classroom and clinical units to nursing students. Includes both teachers primarily engaged in teaching and those who do a combination of teaching and research.
Education Teachers, Postsecondary	0.88	1	Teach courses pertaining to education, such as counseling, curriculum, guidance, instruction, teacher education, and teaching English as a second language. Includes both teachers primarily engaged in teaching and those who do a combination of teaching and research.
Library Science Teachers, Postsecondary	0.84	1	Teach courses in library science. Includes both teachers primarily engaged in teaching and those who do a combination of teaching and research.
Criminal Justice and Law Enforcement Teachers, Postsecondary	0.79	1	Teach courses in criminal justice, corrections, and law enforcement administration. Includes both teachers primarily engaged in teaching and those who do a combination of teaching and research.
Real Estate Sales Agents	0.58	1	Rent, buy, or sell property for clients. Perform duties such as study property listings, interview prospective clients, accompany clients to property site, discuss conditions of sale, and draw up real estate contracts. Includes agents who represent buyer.

Table 6. Occupations did not come up as Analytical which I expect to be Analytical

Job	Probability	Predicted	Description
Compliance Officers	0.47	0	Examine, evaluate, and investigate eligibility for or conformity with laws and regulations governing contract compliance of licenses and permits, and perform other compliance and enforcement inspection and analysis activities not classified elsewhere.

4. Future work and limitations

4.1. limitations

- In this project there is limited amount of labeled training data, having more labeled data set help the model learn more about the data, which results in more accurate predictions.
- In handballing the data, it would have been better if there were a group of people (Delphi method) to label the data to reduce subjective bias.
- Also, it might be good to use text mining techniques for labeling the data in larger scale and faster manner, and then check by human, this way we will reduce both subjective and objective biases.

4.1. If I go back

- when I chose the best model, I will train it for the for whole labeled data set once again, not just splatted dataset, so I have a better trained model.
- In preprocessing the data, I aggregated data feature levels on mean, however I will try median as well since median is robust to outliers.
- I will try random forest as well, and check for calibration of its probabilities due to reasons mentioned in modeling section.
- I will work on feature selections and feature engineering, by using features which have more weight on the model.
- I will search for more resources that give more precise definition of an analytical model.
- I would play with trash hold looking at ROC curve to pick the best threshold based on the importance of precision and recall in prediction of class of interest and project's needs.