Predicting Analytical Occupations

Pak Shing Ho

Approach - Labels

- Hand-labelled 45 analytical and 45 non-analytical out of 967 occupations about which I am most confident.

- Label based on eyeballing the O*NET titles, tasks and job description of each occupation.

- E.g., Analytical occupation title contains ‘Analysts’, ‘Scientists’ and ‘Engineers’. And tasks and job description contain ‘analyze’, ‘analysis’, and ‘research’; Whereas non-analytical title contains ‘Drivers’, ‘Cleaners’, ‘Cooks’, ‘Workers’, etc.

Approach - Features

15 variables (in red) describing level of skills, abilities, work activities, and work styles.

- 4 Skills:
    - Critical Thinking: Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems.
    - Active Learning: Understanding the implications of new information for both current and future problem-solving and decision-making.
    - Complex Problem Solving: Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions.
    - Judgment and Decision Making: Considering the relative costs and benefits of potential actions to choose the most appropriate one.

- 4 Abilities
    - Idea Generation and Reasoning Abilities
        - Deductive Reasoning: The ability to apply general rules to specific problems to produce answers that make sense.
        - Inductive Reasoning: The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events).
        - 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).
    - Quantitative Abilities
        - Mathematical Reasoning: The ability to choose the right mathematical methods or formulas to solve a problem.

- 4 Work Activities:
  - Information and Data Processing:
    - Analyzing Data or Information: Identifying the underlying principles, reasons, or facts of information by breaking down information or data into separate parts.
  - Reasoning and Decision Making:
    - Making Decisions and Solving Problems: Analyzing information and evaluating results to choose the best solution and solve problems.
    - Thinking Creatively: Developing, designing, or creating new applications, ideas, relationships, systems, or products, including artistic contributions.
  - Performing Physical and Manual Work Activities:
    - Performing General Physical Activities: Performing physical activities that require considerable use of your arms and legs and moving your whole body, such as climbing, lifting, balancing, walking, stooping, and handling of materials.

- 3 Work Styles:
    - Conscientiousness:
        - Attention to Detail: Job requires being careful about detail and thorough in completing work tasks.
- Practical Intelligence:
    - Innovation: Job requires creativity and alternative thinking to develop new ideas for and answers to work-related problems.
    - Analytical Thinking: Job requires analyzing information and using logic to address work-related issues and problems.

Approach - Models

Trained 13 different models

1. Decision Tree
2. Logistic Regression
3. Linear Discriminant Analysis
4. K-Nearest Neighbors
5. Gaussian Process Classifier (RBF Kernel)
6. Support Vector Machine (Linear Kernel)
7. Support Vector Machine (RBF Kernel)
8. Naive Bayes (Gaussian)
9. Random Forest
10. AdaBoost
11. Neural Network
12. Voting Classifiers (Hard. Ensemble using model 2 to 9, 4 parametric's and 4 non-parametric's, 3 linear's and 5 non-linear's)
13. Voting Classifiers (Soft. Ensemble using model 2 to 9, 4 parametric's and 4 non-parametric's, 3 linear's and 5 non-linear's)

Why Different Models?

- Small training size with labels unevenly assigned across analytical occupation spectrum. Expect no single model has robust results.
- Eventually, allow using Voting Classifiers to combine models and use a majority vote (hard vote) or the average predicted probabilities (soft vote) to predict the class labels. Such a classifier can be useful to balance out their individual weaknesses.

Advantage of Different Models

- Simple decision tree is the worst because there are only two labels with multiple features. It will make decision by one random feature only, that's why we need Random Forest or AdaBoost (ensemble of simple decision tree).
- Linear discriminant analysis has advantages over Logistic Regression:
    - When the classes are well-separated, the estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem.
    - If training size is small and the distribution of the predictors is approximately normal in each of the classes, the LDA is again more stable.
- K-Nearest Neighbors is a non-parametric approach.
    - No assumptions are made about the shape of the decision boundary. Expect this approach to dominate LDA and logistic regression if the decision boundary is highly non-linear.
- Gaussian Process Classifier (RBF Kernel), allows non-linear decision boundary.
- SVM is good with small training size, take cares of outliers better. RBF Kernel allows non-linear decision boundary.
- Naive Bayes (Gaussian) works well with small training size, but conditional independence (i.e. all input features are independent from one another) assumption rarely holds true.
- Random Forest and AdaBoost: No assumptions on distribution of data. Handles colinearity better than LR, May not be good with small training data size.
- Neural Network: Not good when training data is small.
- Voting Classifiers: Combines and weights different models to balance out their individual weaknesses.

Evaluation

- When occupations ranked by predicted probabilities of any model, the predicted label classes of all models
  - generally agree more the closer to the two extreme ranks.
  - start disagreeing with each other the closer to the middle ranks.
- This result shows that an individual model cannot predict well given observations close to the decision boundary.
  - Likely caused by small training sample, and training data points are not chosen evenly across the analytical spectrum because I labeled occupations about which I am most confident. This causes the vagueness around the decision boundary. And indeed, in reality, it is vague to tell whether a job is analytical or not across the occupation spectrum.

Evaluation

- What occupations did you expect to be analytical and didn’t come up?
    - For example, the following analytical occupations in logistic regression prediction disagrees with some other models: ‘Police Detectives’, ‘Criminal Investigators and Special Agents’, ‘Coroners’, ‘Reporters and Correspondents’, ‘Fire Investigators’, ‘Historians', and some titles contains ‘Inspectors’, ‘Technicians’ and ‘Speicialists’.
- What occupations came up as analytical that surprised you?
    - For example, the following analytical occupations in logistic regression prediction disagrees with some other models: ‘Interviewers, Except Eligibility and Loan’, ‘Travel Agents’, ‘Order Clerks’, ‘Statement Clerks’, and most of the titles contains ‘Clerks’.
- Disagreement across models:
    - LR, LDA predicted non-analytical, all other models predicted analytical: ’Poets, Lyricists and Creative Writers’
    - Most disagreement among different methods: ‘Tellers', ‘Graduate Teaching Assistants’, ‘Travel Agents’, ‘Graphic Designers’, ‘Pilots, Ship', ‘Bookkeeping, Accounting, and Auditing Clerks’.
- The logistic regression model may be bad because the features can be highly correlated, meaning high collinearity.
- No one single model has no unexpected cases.
- Voting Classifier fixed most of the unexpected cases but it is still vague in those cases close to the decision boundary of many models, i.e. the middle ranked cases.

Evaluation

- What other sources of data could be included to make the model more accurate?
  - One can use natural language processing (NLP) techniques to identify analytical occupation. To do so, one can rely on or web-scrape large external text datasets of job description and tasks text data to label occupation classes. Then, use these big text data to train and validate a recurrent neural network model. Finally, feed O*NET job tasks and description text data into the model to predict label classes. One can even combine both text data with numeric data in O*NET to predict label classes.

Assumptions and Limitations

- Assumption 1: Training size of 90, i.e. 10% of the data can represents the population.
- Assumption 2: The 15 features chosen has enough predictability.
- Limitation 1: Training data points are hard to be chosen evenly across the analytical spectrum because I labeled occupations about which I am most confident. This causes the vagueness around the decision boundary.
- Limitation 2: Since the labelled set is small, one cannot perform cross-validation, or it is not helpful. And no regularization can be applied.

- Improvement 1: The first best is to increase the labelled set so that one can perform cross-validation. And so that one can include more features and use regularization.
- Improvement 2: Since the training sample size is small, it's better to label occupation evenly across the analytical occupation spectrum. This can be hard and is a limitation since some occupations are vague in whether they are analytical or not.
- Improvement 3: Some models can perform badly when the features are highly correlated. One can examine the correlation between variables and combine some of them as one by averaging. For example, one can take average of 'Inductive Reasoning', 'Deductive Reasoning', and 'Mathematical Reasoning'.
- Improvement 4: One can use natural language processing (NLP) techniques to identify analytical occupation. To do so, one can rely on or web-scrape large external text datasets of job description and tasks text data to label occupation classes. Then, use these big text data to train and validate a recurrent neural network model. Finally, feed O*NET job tasks and description text data into the model to predict label classes. One can even combine both text data with numeric data in O*NET to predict label classes.