Berkeley Innovation Index
Final Report

Group Members:
My Dinh
Jessica Gu
Aaron Lu
Dayou Wang
Yan Zeng
Yujun Zou

Course: Data-X Fall 2017
Due Date: December 8th, 2017

I. Introduction:

Berkeley Innovation Index is a research project developed by Sutardja Center for Entrepreneurship and Technology as an assessment tool to evaluate innovation mindset. This research is based on a survey of 36 questions to measure six different key traits of innovative individuals: Trust, Resilience, Diversity, Collaboration, Belief, and Effectiveness. The analysis of innovation mindset can be an essential tool for HRs to access the candidates' personality and behavior besides their resumes and interviews response.

Applying machine learning techniques, our group creates a simple yet powerful way to measure candidate's personality and employability. The final model comprises both unsupervised learning algorithms to cluster individuals into four different categories and assign class to new fill-out survey users.

This model not only aims to provide a comprehensive picture about candidates but also helps HR select the best fit who carries the traits for their team's need. Moreover, by classifying the individual into personality group, this tool can be a valuable source for candidates to access their strengths and weaknesses, thus help them grow and develop certain abilities.

II. Data:

The dataset contains responses to 36 questions from over 2000 test takers through the survey in the website: https://berkeleyinnovationindex.org/. Out of 36 questions, 29 questions are designed to measure personality traits and 6 questions ask for the demographic information of the test takers. The first 29 questions are based on a 5-point likert scale, with 1 marked as "Strong Disagree" and 5 as "Strong Agree". The response thus can be regarded as ordinal data. Nominal data such as Stage, Study, Gender, Age contains 3, 3, 2 and 3 levels respectively. Area and Country are also categorical responses, but with unlimited levels.

III. Methodology:

1. Data Preprocessing:

Based on our quick observation of the dataset, we find out that some responses are either uniformly marked as 1 or 5. We decide that they are test data from the developers and thus are removed from our working dataset. Based on suggestion in Ojala's paper, the team also reverses the response of "QT2", "QT4", "QT5", "QP1", "QP3" to account for the opposite framing in the questions.

Secondly, the team transforms the non-numerical data to numerical ones for features like "Area" and "Country". To further explore the pattern of our data, new features such as "overall_score" which sums up the total score for each user, "extreme_score" which counts the number of times each user picked "5" or "1", and the "neutral_score" which counts the number of times each user picked "3" are added from this state. We later perform PCA on this dataset to observe any clustering pattern. Heat maps are also utilized to check the correlations between variables and to remove the highly correlated ones.

2. **Data transformation and feature engineering:**

We take multiple approaches to transform our data into a format where we can build our machine learning models on. Because of the nature of our data, instead of performing a standard normalization with mean and standard deviation, we decide to scale the response into values between 0 and 1. When fitting in the supervised clustering model, we use SMOTE(synthetic minority over-sampling technique) to account for the imbalanced data (because one out of five outputs has as many as seven times amount of data than the rest). We also perform PCA and Recursive Feature Elimination to reduce the dimensionality of the feature space from 36 to 33 and normalize the data set before the team trains different models.

Beside from scaling the response into values between 0 and 1, we also observe that though the data is in numerical value, the first 26 questions have a ranking structure while the rest of 6 questions have nominal structure, which are essentially two types of categorical data.One of the problem with this dataset is that, the difference between each marked scale is uniformly. For example the difference between response 2 and 3 (disagree vs neutral), and the difference between response 3 and 4 (neutral vs agree) are the same. Thus, the difference does not reflect the true dissimilarity between responses. This could be a problem when performing clustering algorithm because most of the techniques are based on distance or similarity metrics.

In order to account for this problem, we decide to perform NonLinear PCA (NLPCA) to transform categorical variables into numerical values through optimal scaling process. The solution is attained such that the loss function, which measures the departure from homogeneity is minimized. We check transformation for each variable and conclude that the solution obtained by NLPCA is reasonable.

3. **Model Building:**

A. Supervised Algorithms:

Our first approach is to predict whether this test taker succeeds in his/her career based on their response to the questionnaire. We train the following six supervised clustering algorithm: logistic regression, support vector machine, perceptron, KNN, XGBoost and random forest. The target is the feature “ER” (been successful in innovation) with five levels (1,2,3,4,5). After the comparison of each testing error, the team picks random forest as the best supervised clustering algorithm to treat our data (training accuracy 100%, testing accuracy 73.73%). The accuracy score improves the random guess rate from the base model by almost 40%. Also, it is a good score since the target has five levels.

B. Unsupervised Algorithms:

Though our first supervised model to predict whether the candidates succeed in their career attains reasonable accuracy, this approach does not satisfy our main goal to provide insights about the hiring candidates. Thus we also try unsupervised algorithms to divide individuals into clusters based on their responses, hoping to attain distinct traits for each cluster group.

Since we have two preprocessing datasets, we perform different clustering algorithm on both data.

Data type	Model	Assessment/Problem
discrete	K-modes	Aftering feeding in the normalized data sets, the team finds out that the “similarity” score is hard to interpret for the generated clusterings and hard to display the “distance” concept.
discrete	Affinity propagation	The output number of clusters is over 100, which is not suitable for the future interpretation
continuous	K-means	Quantify ordinal values into numerical value; Over 95% fit into one single cluster, which may not cluster completely.
continuous	Hierarchical	
discrete	K-means	Feed the model with normalized data sets. Use silhouette score to pick the optimal number of cluster (n=4), which generates four clusters with relatively balanced weight which suitable for future interpretation.

IV. Result and Conclusion

1. Final Model:

After attaining group clusters and labels, we decide to classify a new user response into one of the four groups. The similar techniques discussed in the previous sections are also employed in this stage to attain a good high accuracy score. After revisiting the result from PCA in EDA stage and the dendrogram on NLPCA transformation, we conclude that the cluster of minority groups is the result of either outliers or distinct characteristics group. Though K-Means on normalized dataset provides a balanced clustering, NLPCA transformed dataset seems to have the model pick up the irregular traits. However, because of the time constraint, we decide not to complete outlier detection analysis with this data at this time. We thus only incorporate the clustering result obtained by K-Means on normalized data into our final web application.

Based on the statistics produced by K-means model, the team performs analysis on each cluster and comes up with the following group description:

- Administrator: Performs well especially in routine works and might need more incentives for potential challenges. Possible Occupation is office administrators.

- Captain: Team leaders, always optimistic and self-determinate about challenges, and good at communicating with teammates. Possible Occupation is senior management.

- Critic: Curious people that are self-motivated for challenges and attempts for every possible outcome. Possible Occupations are researchers and supervisors.

- Communicator: Good at delivering and sharing thoughts with diverse groups of people, and are potentially leaders with proper training. Possible Occupations are public relations specialists and HR.

2. User Interface:

The team uses django framework deployed on heroku to build a web-based user interface: https://bii.herokuapp.com/bii/.

The user interface has two pages. The first page is the survey page where users can enter their response. After clicking on the “submit” button, the response will transfer to the backend and the algorithm runs. Then second page will pop up and display the results of two parts: one is the innovation index score, the other one is the probability of assignment and the description of the four groups.

This user interface is valuable to both HR and individual candidate. For instance, HR could use it as a tool to compose an ideal team to boost team performance. Individual candidates can use this tool to evaluate and improve themselves based on the strength and weakness report.

V. Future Improvement

1. Survey Question Revision

The team could add validation questions to determine whether the data is ‘useful’. Useless data, such as random answers, should be excluded from data training. We could also add critical questions that have more weights towards the final clusters.

2. Unsupervised Continuous Clustering Models Research and Analysis

The team would like to further implement different clustering on both dataset to improve the interpretability of each cluster. In the case of NLPCA transformed data, we will perform another clustering algorithm on majority group to get more reasonable results.

3. Model Update: Dynamic Model

The team will change the current static model to dynamic one, and update the parameters of the clusters (e.g. once a week) when more responses are added.

VI. Team Member Contribution

My Dinh: EDA, nonlinear PCA transformation, unsupervised clustering model building, interpretation of hierarchical clustering result.

Jessica Gu: data preprocessing and survey analysis, analyzes the intrinsic relations within the cluster and differences between clusters, affinity propagation clustering algorithm research.

Aaron Lu: data preprocessing, unsupervised clustering model building and model deployment, user interface development

Dayou Wang: analyzes the intrinsic relations within the cluster and differences between clusters

Yan Zeng: data cleaning and feature engineering, supervised and unsupervised clustering model building

Yujun Zou: supervised clustering model building, user interface development

VII. Mentor Feedbacks

The group has been working with mentor Alexander Fred Ojala on a bi-weekly basis to discuss the questions the team has in mind and to receive guidances on the approaches the team could take for next step. When the team looks back at the progress, the team finds it very important to clarify the problem statement in the initial phase and, thanks to mentor Alex who has spent much time with us weekly out of his busy schedule to provide helpful suggestions. The insights the team gained at the early stage of the project set as a clear objective to move forward later on.