Optimizing Workplace Productivity

IST 718: Big Data Analytics

Dorothy Fang, Courtney Hrdy, Shannon Gambuti

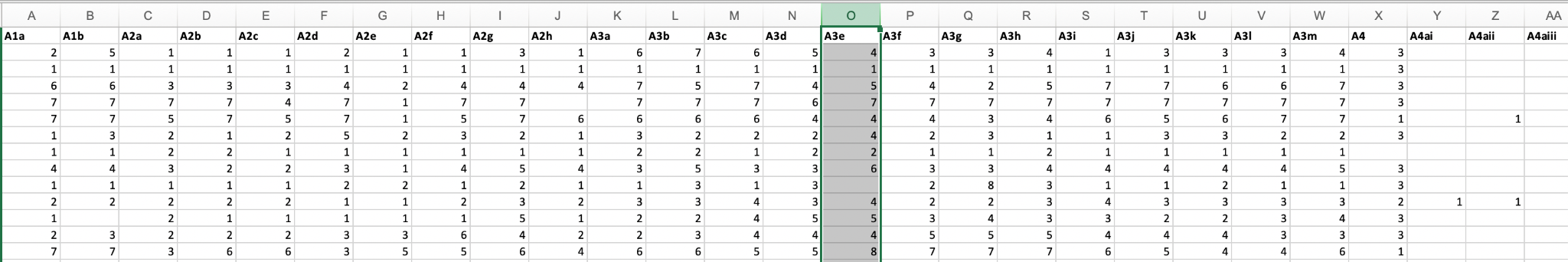
December 17, 2021

**Introduction**

Our project aims to explore the topic of human resources analytics, specifically on how employees interact with their workplace. There are many existing projects currently in the topic that explore employee attrition, employee absenteeism, employee retention and how financial incentive motivates employees’ productivity. However, there is a lack of projects that offer a more holistic view and are people-centric in terms of workplace evaluation. For example, looking at how overall happiness or satisfaction of employees can directly link to increases in productivity. The objective for this project is to evaluate and identify key areas of a workplace, beyond financial incentive, that can increase employee productivity. Productivity is the main driver of profitability since employees are those working towards key performance indicators. By understanding how these key areas boost productivity, human resources departments at enterprises can make better internal investment decisions. In our model, we are expecting to see indirective incentives that improve the work environment such as opportunity for position advancement, advocacy for diversity and inclusion, offering of mentorship by senior leaders, better accommodations, etc. to have a positive impact on employee productivity. With COVID-19 and how the workforce is more attracted to workplaces that are more accommodating and embracing of diverse backgrounds, enterprises would have to lean into these insights to attract and retain their prospective and current employees. Our project was completed in Jupyter Notebook using Pyspark.

**Dataset**

To accomplish our project, we used survey data from the Public Sector Commission of Western Australia. The survey’s objective is to bring light to feedback on how to enhance integrity, effectiveness and efficiency to the public sections.The survey was completed by employees of 11 public sector organizations. In 2016, there were a total of 3883 valid responses. The questionnaire has 109 questions. These questions act as the variables in the dataset which are all categorical ordinal variables. Most questions are structured on a scale of 1-8 corresponding to responses between strongly agree - strongly disagree and very satisfied - very dissatisfied. There are other questions with responses between never - very frequently or yes/no. Below is a snippet of what the dataset looks like:

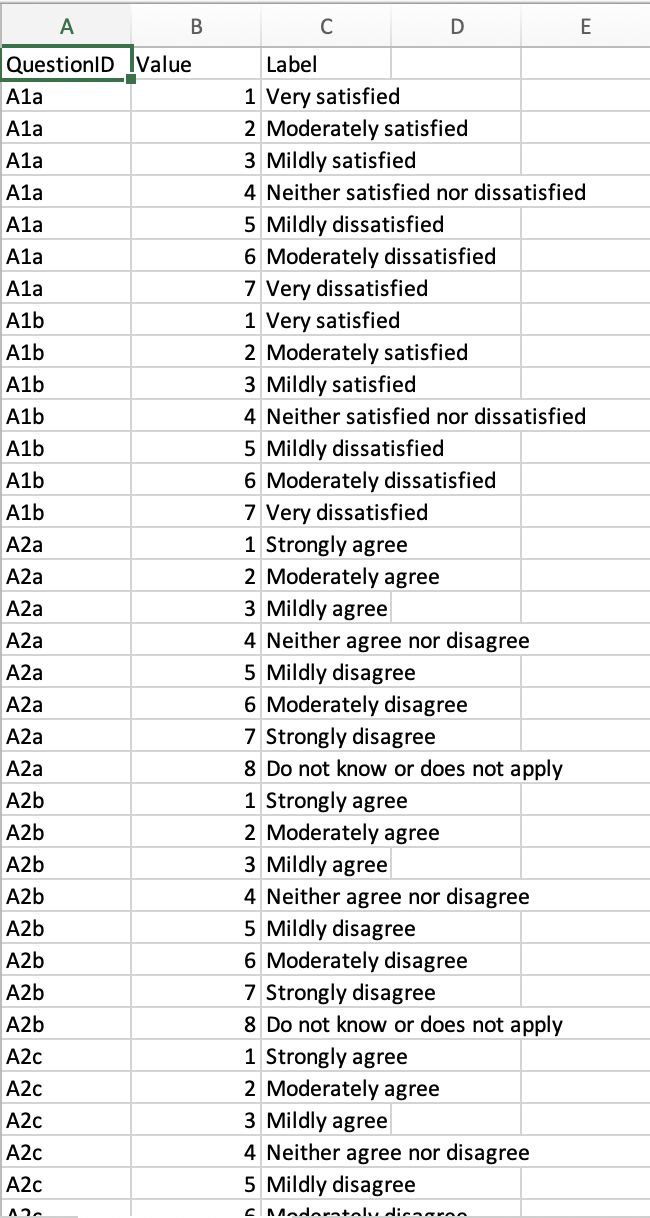


The questions' responses are encoded. The dataset also comes with a question key and a response key that provide clues to what the encoding means.

Below is a snapshot of the question key. It has the actual questions along with the question ID that correspond to the variables in the actual dataset.



Below is a snapshot of the response key. It has the question ID, along with what the values in the dataset mean in terms of its categories.

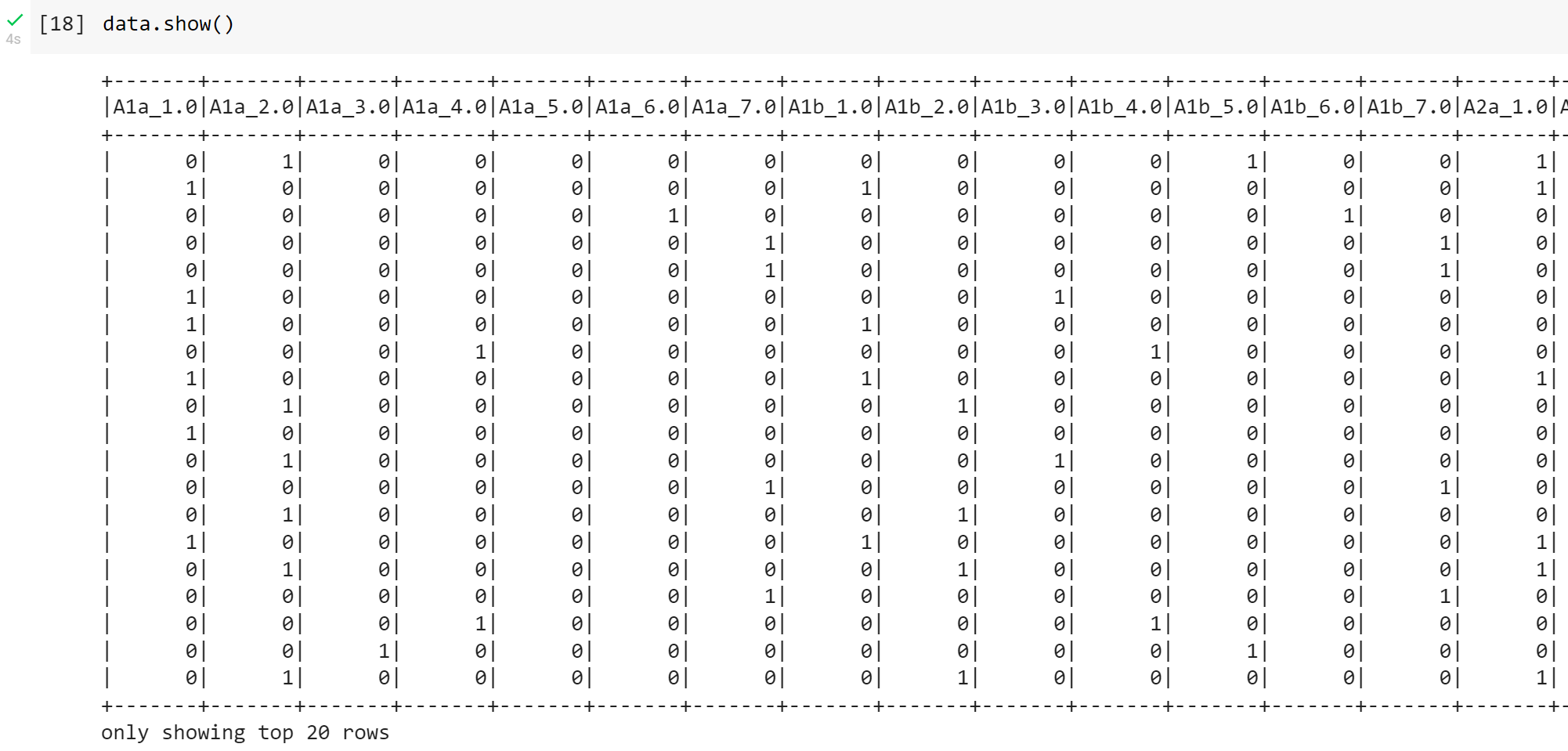


**Project Objective:**

Our objective is to find what other factors in the workplace are correlated to high productivity. Thus, we look to questions B3b in the survey for some answers. Question B3b asks the employee to answer their agreement to the statement, “My work group achieves a high level of productivity.” The responses range from strongly agree to strongly disagree. If someone answers 1 for question B3b, they strongly agree with the statement and if someone answers 8, they strongly disagree with the statement. Thus, the objective is to identify the people that answered 1 for question B3b and find out their answer to other questions.

**Data Cleaning**

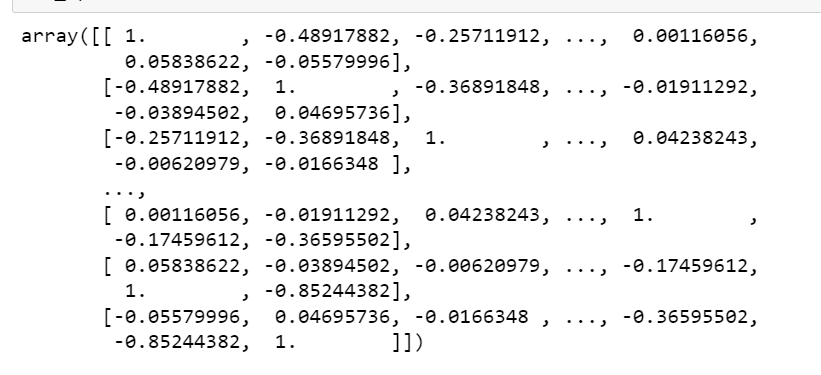
Prior to our analysis, we created dummy variables for each question. The rationale behind this is to easily identify who answered 1 for question B3b. After this transformation our dataset contained either a 0 or 1 value for each response for each question.



**Exploratory Analysis (Unsupervised learning)**

**Correlation Analysis**

After creating dummy variables for each question response, our data frame contained 559 columns. We performed a correlation matrix as seen below.



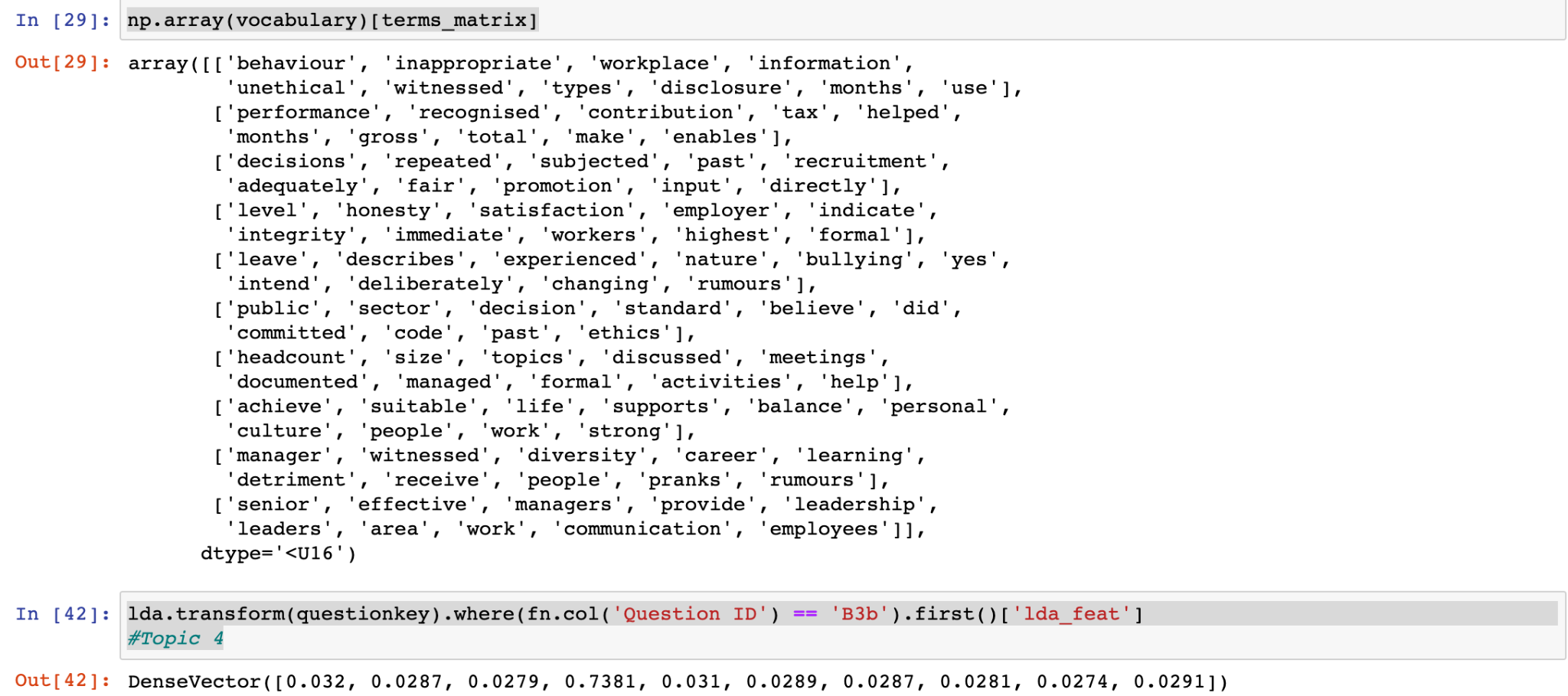
The highest correlation value was .44, telling us that our data is not highly correlated.

Looking at the Agency Size attribute, we found that 63% of the responses were made by employees at large agencies with over 1,000 employees. Roughly 30% of responses were recorded from employees at agencies with 200-1,000 employees and 7% with less than 200 employees. Our target variable is B3b\_1. This variable corresponds to the response, ‘Strongly Agree’, for the question: My work group achieves a high level of productivity. The majority of employees responded with strong agreement and agreement.

**LDA**

Since there were 109 questions present in the survey, we wanted to see if we can find some commonality between questions and be able to classify them under themes. Using Latent Dirichlet Allocation (LDA) topic modeling on the question key dataset, we wanted to see how similar each question is to each other. The modeling was done on the question text, therefore, we had to preprocess the data by removing punctuations and stopwords. Then, tokenizing, normalizing and standardizing the term frequency inverse document frequency (tf-idf) of each question. Using the processed tf-idf, we look for 10 topics within the 109 questions. The limitation with LDA is that the topic naming is human interpreted from the top keywords of each topic. Since LDA does not have a golden standard for telling how many topics to select for the best telling answers, we chose ten because the keywords made the most sense.

Below are the top 10 keywords for ten topics:



Based on the keywords, we interpreted the topics to be:

* Topic 1: Ethical Behaviors of employee/senior management
* Topic 2: Recognition of achievement
* Topic 3: Hiring decisions
* Topic 4: Integrity and Satisfaction of employee
* Topic 5: Bullying at the workplace
* Topic 6: Code of Ethics
* Topic 7: (Inconclusive)
* Topic 8: Work-life balance support
* Topic 9: Diversity in management
* Topic 10: Senior management leadership and communication.

Question B3b, belonged to topic 4 which is related to the integrity and satisfaction of employees.

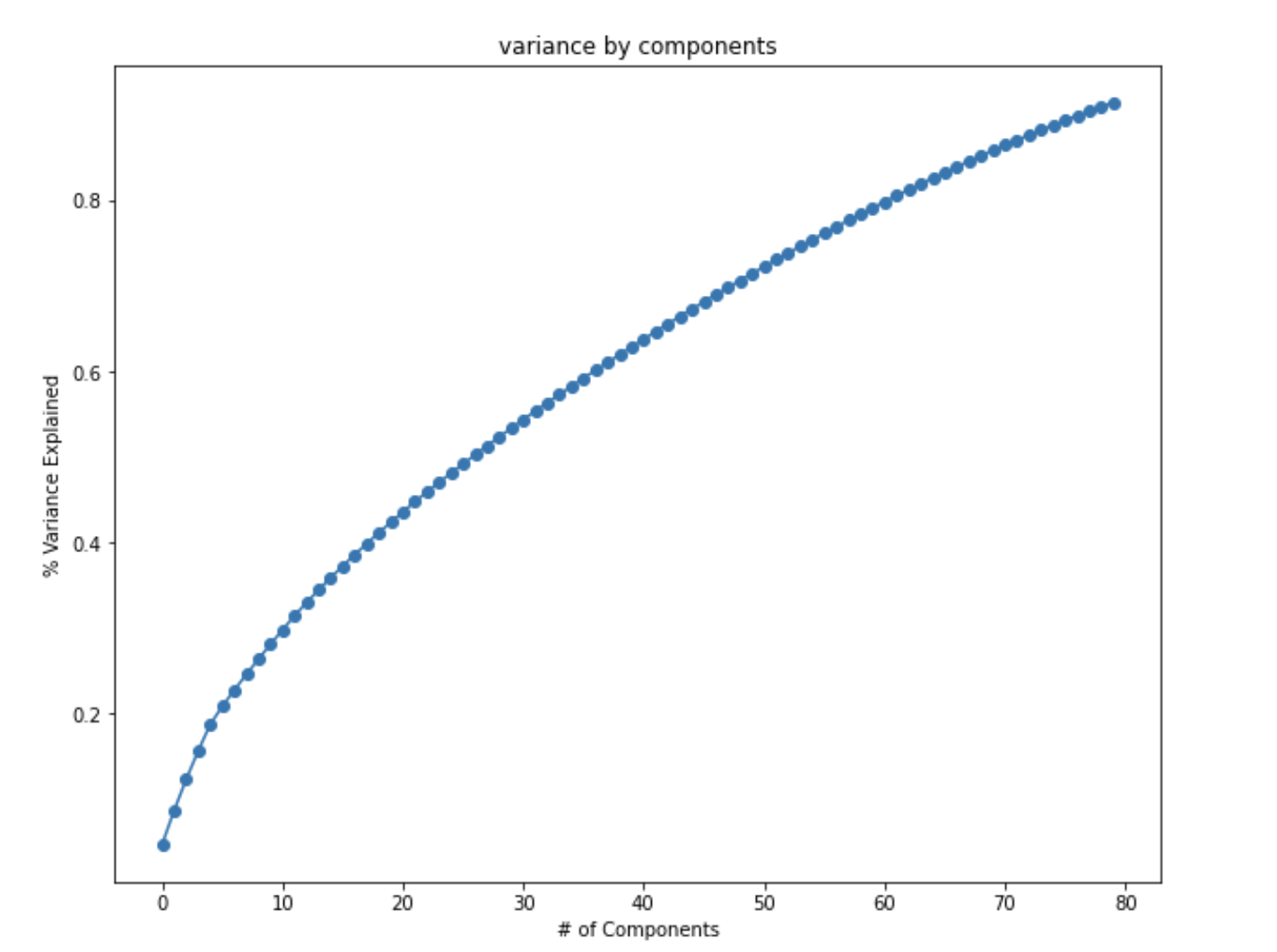
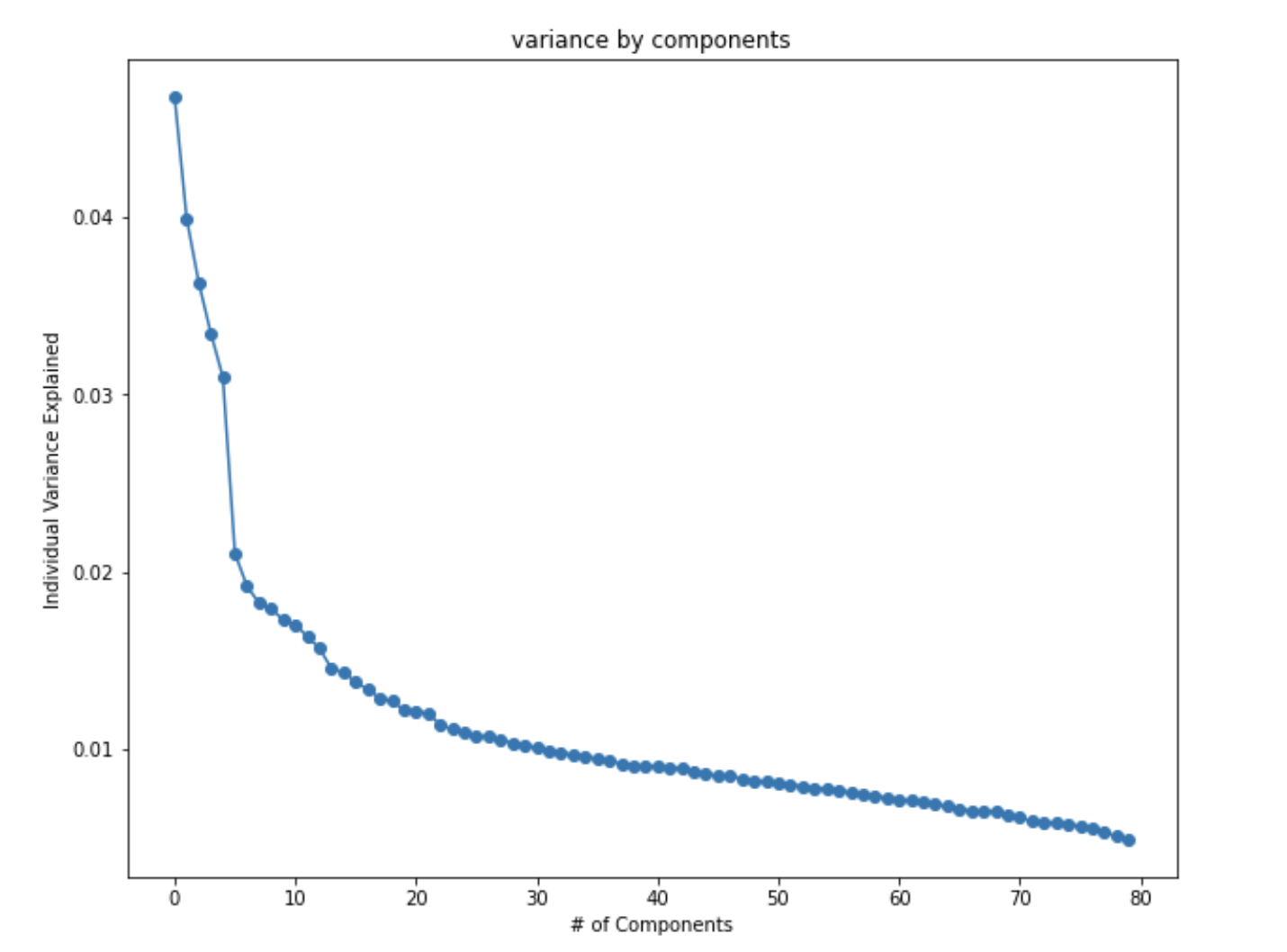
Some other questions in topic 4 include:

* A1a: Please indicate your level of satisfaction with: My job overall
* A1b: Please indicate your level of satisfaction with: My agency as an employer
* A2d: I have the authority (e.g. the necessary delegations, autonomy, level of responsibility) to do my job effectively
* A3e: Recruitment and promotion decisions in my agency are fair
* A3h: My agency is committed to health and wellbeing within the workplace
* B1b:​​ My input is adequately sought and considered about decisions that directly affect me

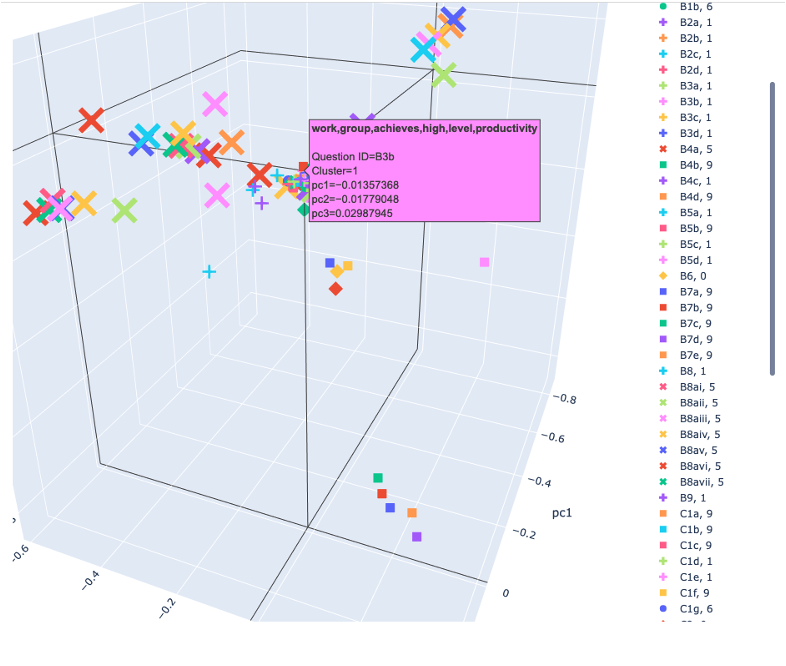
This list is non-exhaustive, but we can see that satisfaction and work autonomy questions are all a part of the same type of question with question B3b.

**Principal Component Analysis and KMeans**

For exploratory data analysis of the questionnaire, the normalized features of the filtered question words from the tf-idf matrix were mapped into a low dimensional space using a Principal Component Analysis (PCA). The variance explained ratio was used to understand how the variance is explained by each principal component, with the first principal component having the highest variance from the normalized features. The decrease in the variance explained for each principal component is used to understand variance in the rest of the dataset. It took 80 principal components to understand 90% of variance in the dataset, indicating that all of the questions are very similar.



Kmeans uses the normalized tf-idf features from the PCA pipeline and outputs data into nine clusters. The silhouette score is the measurement used to decide the number of clusters. Silhouette Score measures the average similarity of the objects within a cluster and that distance to other objects in other clusters. The beauty of having the PCA is the ability to reduce the number of variables by combining them into meaningful features where they are orthogonal to each other. The questions clustered with the target variable would be expected to have high feature importance or correlation in the answers, which they do. The clustering identifies questions associated with personal perceptions of the workplace such as “I feel a strong personal attachment to my agency” and “I am proud to tell others I work for my agency.” This establishes that it is important for companies to build and foster a culture of belonging for their workforce.

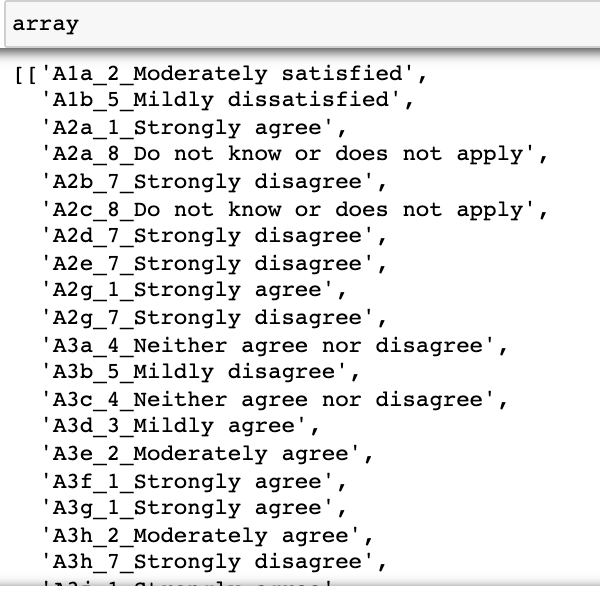


[3D PCA](https://chart-studio.plotly.com/~courtneyhrdy/7/#/)

**Association Rules Mining**

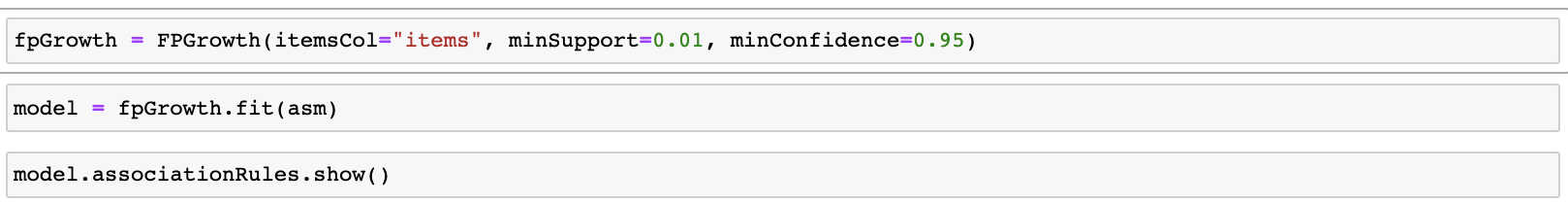
Association Rules Mining is typically used in market basket analysis, where researchers try to identify what is often bought together in a transaction. An example question could be, given someone purchased diapers, what is the probability that they also purchase eggs? The idea to use association rules mining in this case is to treat each response like a transaction. Therefore, what we want to find out is that given someone answered B3b\_Strongly Agree, what is the probability of other responses to other questions.

This task required a lot of data preprocessing. We were able to manipulate all the responses into the a list, like below to show what response the respondent gave to each question:





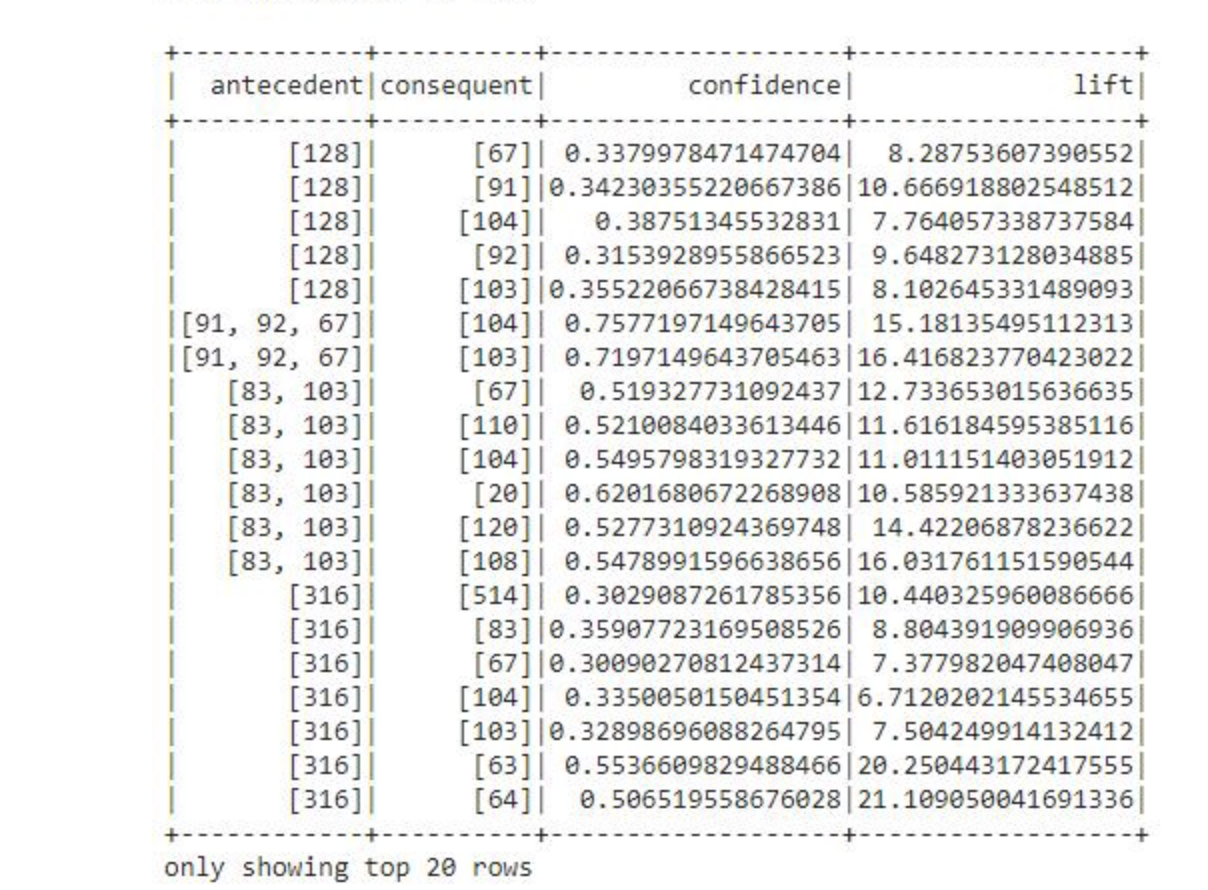
We planned to use the fp growth algorithm to conduct the association rules mining.



However, given the size of our data, we ran into some computational limitations and were unable to run the result.

We were trying to achieve a result like the screenshot below. It was obtained from <https://towardsdatascience.com/market-basket-analysis-using-pysparks-fpgrowth-55c37ebd95c0>.

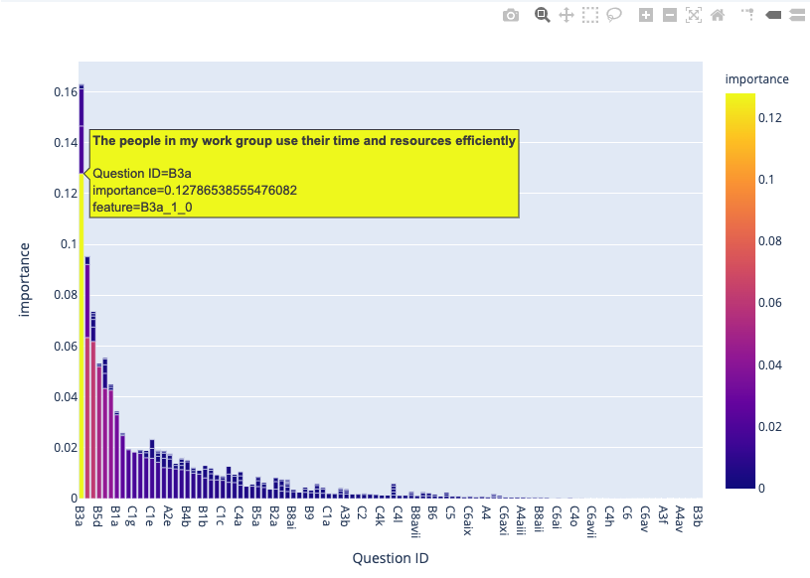
Then to find that given the consequent being B3B\_1\_0, what are the antecedents based on the lift since it is an indicator of how correlated the antecedent is to the consequent.



**Supervised Learning**

**Random Forest**

The Random Forest is used to look at features by their mean decrease in impurity to evaluate their importance in classification. Random forest removes correlation by fitting decision trees to subsets of the features and is supposed to be a good way to assess feature importance without overfitting the data. Forest models are an ensemble of decision trees, each one able to predict its own response to a set of input variables. Each tree is created on a different sample selected with replacement, so the different combinations of cases and inputs for the splits are more varied. Results are combined to provide the final prediction, in which the strongest association with the target is used in the splitting rule from all available inputs. Machine learning applications to survey research have recently been used to estimate response propensities and are able to process given responses to make predictions. Survey answers that have the highest feature importance in relation to the target variable (question b3b) had responded with being very satisfied to the following questions: “The people in my workplace use their time and resources efficiently”, “The people in my workplace are committed to providing excellent customer service and making a positive impact in the community”, “In the last 12 months my company has implemented innovative processes and policies”, and “Your coworkers treat employees from all diversity groups with equal respect.”



[Link to Random Forest Feature Importances](https://chart-studio.plotly.com/~courtneyhrdy/4/#/)

| Accuracy | Precision | Recall | F1 Score |
| --- | --- | --- | --- |
| 0.81 | 0.85 | 0.70 | 0.77 |

## From our results, we see that the Random Forest has fewer false positives than it does false negatives, meaning each tree can capture impurity but it is overfitting the data. However, evaluation of model performance transcends looking at the traditional metrics, as we are predicting a nuance of human behavior. The algorithm is making predictions using survey answers without understanding the sentiment of the question. In contrast, our unsupervised learning models used the transformed TF-IDF features of the questions without learning the correlation of responses. To holistically evaluate human factors like productivity, the combination of unsupervised and supervised learning methods needs to evaluate interaction sequences that reveal cognitive and emotional elements of the predictions to bridge the gap between computational algorithms and human reasoning. So, if we are only going to make decisions from this dataset using the basic metrics (accuracy, precision, recall, f1 score) we are missing out on understanding user responses.

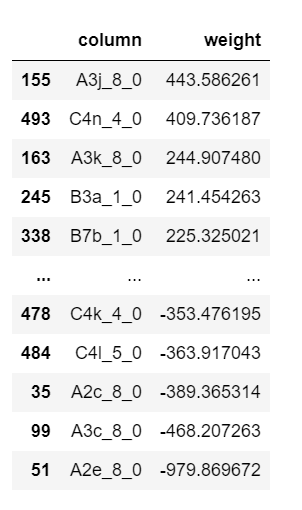
**Logistic Regression**

We created three logistic regression models: one using all features as the independent variables, one using the features with highest importance determined by Random Forest, and last using all features against the dependent variable B3b\_7, indicating ‘strongly disagree’ to our target question. To measure the accuracy of our logistic regression models with finding the Area Under the ROC curve. This AUC value tells us how good our model is at distinguishing between classes. All three of the models had negative intercepts. The negative intercept corresponds to the estimated probability of the response being less than 50% when all model covariance equal 0. Below are the AUC values for each of the three models.

| All Features | Random Forest Features | All Features - B3b\_7 |
| --- | --- | --- |
| 0.81 | 0.92 | 0.47 |

As you can see, the model predicting B3b\_7 performed the worst. We believe this is due to the low sample size, given that not many employees responded with strongly disagreement for our target question. Below are the weights associated with each independent variable determined by the Logistic Regression model. The columns with higher weights indicate higher importance within the mode

**All features predicting B3b\_1:**

.

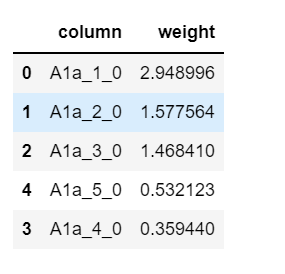
When all features were used as the independent variables, the following questions and responses had the highest weight:

* A3j: I feel a strong personal attachment to my agency
  + Do not know or does not apply
* C4n: Have you witnessed any of the following types of unethical behavior in your workplace in the last 12 months, and if so how often?: Engaging in criminal behavior outside work
  + Frequently (7 to 10 occasions)
* A3k: My agency inspires me to do the best in my job
  + Do not know or does not apply
* B3a: The people in my work group use their time and resources efficiently
  + Strongly Agree
* B7b: Do you believe any decision made in your agency in the past 12 months did not comply with a Public Sector Standard: Yes, and I lodged a breach of standard claim?
  + Yes

**Random Forest Features predicting B3b\_1:**

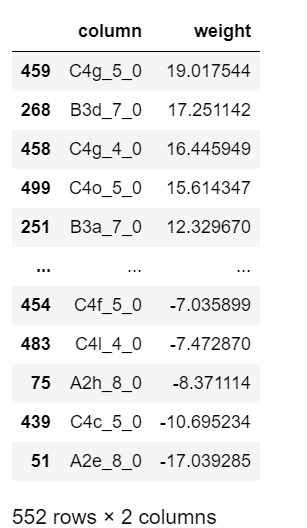
This model had the highest AUC score.

In using logistic regression to predict whether a person answered the 1 in question B3b, the following variables were indicative.



We see that people who answered “Very satisfied” to the statement “Please indicate your level of satisfaction with: My job overall” had the largest weight in predicting whether someone answered “strongly agree” to “My work group achieves a high level of productivity.”

**All Features in predicting B3b\_7:**



The people who answered “strongly disagree” to “My work group achieves a high level of productivity”, are seen to also answer to these question with these responses:

* C4g: Have you witnessed any of the following types of unethical behavior in your workplace in the last 12 months, and if so how often?: Illicit drug use and/or alcohol intoxication.
  + Very frequently (more than 10 occasions)
* B3d:The people in my work group are committed to providing excellent customer service and making a positive difference to the community
  + Strongly disagree
* C4g:Have you witnessed any of the following types of unethical behavior in your workplace in the last 12 months, and if so how often?: Illicit drug use and/or alcohol intoxication.
  + Frequently (7 to 10 occasions)
* C4o:Have you witnessed any of the following types of unethical behaviour in your workplace in the last 12 months, and if so how often?: Secretly holding another job outside government without agency permission
  + Very frequently (more than 10 occasions)
* B3a: The people in my work group use their time and resources efficiently
  + Strongly disagree

**Conclusion**

In conclusion, the questions that were identified in the unsupervised machine learning algorithms were consistent with the feature importance we got from supervised learning. An employee’s perception of their workplace is most predictive of a work group achieving high levels of productivity. Our highest performing models were the Logistic Regression using the features identified with Random Forest and the Random Forest Model.