**Individual Peer Evaluation Form**

Your name: *David Culhane*

Write the name of your classmate you are preparing this review for in the designated column. Using a scale of 1-4 (1=strongly disagree; 2=disagree; 3=agree; 4=strongly agree) answer each question. If you aren’t able to answer the question based on what is posted in the discussion board, reach out to your classmate for more information via the discussion board. Total the numbers in each column. **Make sure to answer the questions on the 2nd page.**

|  |  |
| --- | --- |
| Evaluation Criteria | Peer Name:  Andrew Pfeifer |
| Has plan in place to complete course project. | 4 |
| Has found datasets/data sources to support project idea. | 4 |
| Has solidified project idea. | 4 |
| Has identified resources for project. | 4 |
| Topic is related to data science and demonstrates topics learned to date through program. | 4 |
| Risks and potential issues have been identified. | 4 |
| TOTALS | 24 |

Feedback on Individual’s project topic:

1. How clear is the classmate’s project topic? What questions does their topic make you consider?

*The topic for Andrew’s project is clear – the detection of fraudulent credit card transactions. This topic makes me think about how to identify if a transaction is fraudulent. Using location or other information that would not be anonymous could be helpful, but it would have severe ethical implications. When using anonymized data, the challenge will be how to detect fraud without using location or ethically sensitive data.*

1. What risks or issues should your classmate consider while working on their project?

*A major issue Andrew already identified is the lopsided nature of his dataset. Fraudulent transactions are in a significant minority when compared to legitimate transactions. This imbalance has to be dealt with. While this issue was identified, a solution was not specified. Solutions could include balanced modes in model creation, up-sampling of fraudulent transactions, and down-sampling of legitimate transactions.*

*Model selection is another issue Andrew identified, properly stating that the target variable is classified binarily. Logistic regression will likely be a good choice for the model to use. Random forest classification could be good as well depending on the data. I would be interested in seeing how support vector machines would be helpful in this situation, since my understanding is that they are useful in separating areas in plotted spaces. The multidimensional nature of the data would make it hard to plot in a two-dimensional space. The choice of kernel for the SVM would be important as well, which would be influenced by the spaces fraudulent transactions occupy in the plotted space.*

1. Additional suggestions/comments that might be beneficial to your peer?

*I think having a backup dataset similar and dissimilar to your primary dataset to fall back on is a good idea. It shows foresight in case your primary dataset has some kind of issue. One of Andrew’s backup datasets having a field with information about the distance between the listed and immediately previous transaction is interesting. That is something I would want to have in a dataset.*

Adapted from a peer evaluation form developed at Johns Hopkins University (October, 2006)