1. What is the overall objective of your project?

The overall object of the project is to identify receipts that can link to fraudulently employee activity.

1. Where/how are you going to obtain the data? For example, do you need to scrape the data from a web source, contact a company or individual for access/permission?

Joel’s father is the owner of a chain of natural foods stores. Joel is an employee there and because of this has access to all the receipt data. Any and all permissions are taken care of through Joel.

1. Are there any copyright or terms of service limitations that may affect your ability to work with the data?

No.

1. What type of machine learning algorithm / process do you anticipate using (e.g., classification, regression, clustering) and if applicable, what variable(s), in particular, will you be trying to classify, etc.

We will be using classification types of algorithms. We decided to go this route because there is data that is known to link to fraudulently employee activity. We will be classifying each receipt as potentially fraudulently or not.

1. How many and what type of attributes do you anticipate? (e.g., less than than 10 vs. 50-100, numeric vs. nominal, time-series data, etc.)

There will be a mix of numeric and nominal data. Examples of this are employee names, customer names, return dollar amounts (negative numbers), transaction numbers, lines of the receipt, tenders, transaction locations, transaction start times, and register numbers.

1. How many instances or records do you anticipate?

There is about 3 - 4 years worth of receipt data. However, we will be filtering through most of the receipts that don’t have negative total values to them. In other words, any receipt with no items being returned won’t even be considered. We also have about 100 or so instances that will be used as our training set.

1. What do you see as the biggest risk to the success of your project?

There could be a problem with overfitting, as the number of employees that actually steal money is a lot lower than the total number of employees.