**Project Two – Combining Multiple Classifications**

**Part of Developing a Data Mining Portfolio**

## A data mining Portfolio

Taking a portfolio approach to your data mining projects can provide you with evidence of your data mining abilities and of the types of work you can conduct. This can be valuable for providing information about your technical abilities, your ability to conduct data analytics, and conduct data mining activities. From an employment perspective it can help employers quickly identify who can do the necessary work. This is particularly useful when your data mining projects look like the types of work you would do when hired. With data mining this can be achieved by developing a portfolio of projects. The more realistic the projects are the more assurance you provide that you will be an asset to an organization.

## Your Data Mining Portfolio Contents

One question to address is where to put your projects. One recommended locations is Gitbub. Another is your own log. There you will want to make your code visible and well-documented. This will provide an easy access to your work as well as allow evaluation.

It is useful to document your projects as well as possible. Typically, including a README file explaining how to set up and run the code, discuss needed packages , and discuss the data, possibly including a data dictionary.

There are several different types of projects you will want to include. These will be based on the data mining methodology. Ideally, you would have multiple projects of each type and especially, types based on the type of job you want. For instance,

Jobs that require a lot of machine learning suggest that you provide more end to end projects that use machine learning. On the other hand, for analyst positions data cleaning and storytelling projects may be more relevant.

**End to End Data Mining Project**

The exploratory data cleaning and analysis project provide information about how well you can extract insights and present them to others. However, they don’t demonstrate that you can build systems that are customer-facing. Ultimately, that is the target of the data mining models being developed – deployment for use. Customer-facing systems involve code that can be run multiple times with different pieces of data to generate different outputs. Possibly, downloading new data daily or hourly for subsequent use.

An end to end project takes in and processes data, then generates some output. This involves developing data mining models using machine learning algorithms, but can also be another output, such a database processing.

The key here is to make the system flexible enough to work with new data and also i make the code easy to setup and run. Here are typical steps you’ll need to follow to build a good end to end project:

* Find an interesting topic – You won’t focus on a single static dataset, so you’ll want to find a topic instead such as flights, electricity pricing, health.
* Import and parse multiple datasets
* Create classifications or predictions. Determine the needed features and create training and test data. Using that to make predictions or classifications.
* Clean up and document your code

**Data mining project types**

This assignment is focused on End to End to projects. However, the other types will be briefly described.

**Data Cleaning Project**

A data cleaning project taking a combination of datasets and make sense of them. This is a key step in data mining. This type project involves taking conducting the data preprocessing steps: taking messy data, cleaning it up and conducting a variety of statistical and summary analysis. A data cleaning project demonstrates that you can reason about data, can take data from many sources, and consolidate it into a single dataset.

You’ll want to go from the raw data to a version that’s easy to do analysis with. In order to do this, you’ll need to:

* Find a messy dataset. Avoid picking anything that is already clean — you want there to be multiple data files, and some nuance to the data.
* Pick a question to answer using the data. Explore the data
* Clean up the data
* Do some basic analysis. Try to answer the question you picked initially
* Present your results. Use R markdown to do the data cleaning and analysis

**Data Storytelling Project**

Type of project demonstrates your ability to extract insights from data and persuade others and is closely related to the work in the data cleaning activities. This type of project involves taking a set of data and telling a compelling narrative with it. For example, using the flight delays data set we used you could use the data to show that there are significant delays at certain airports. You can suggest solutions to address such as changing depart times or by changing the routing.

Visualizations can be effective here and judicious use of graphs and text can play a key role. Here are some stepsto build a good data storytelling project:

* Find an interesting dataset
* Explore a few angles in the data. Explore the data. Look for correlations in the data and create charts.
* Write up a compelling narrative. Identify the most interesting angle from your explorations and write extensive analysis of the results of each step, and what they tell a reader
* Present your results. Use R Markdown including documentation so that your code and logic can be followed explaining your process as you can