Daniel Valle

**Capstone project 1 – Data wrangling**

This document describes the main steps undertook to clean the data for the capstone project 1.

Three of the features in the data set contain missing values. They were converted into 0s. Note that these features comprise three data types: string, float and boolean.

The feature corresponding to the 911 emergency address has been changed to string. From this feature is extracted a new feature of type string corresponding to district in which the incident occurred.

From the location feature, two features corresponding to the longitude and latitude are extracted via two functions that extract this information making use of regular expressions. These features are added to the original dataset.

The information contained in the ‘calltime’ and ‘calldate’ columns are transformed into datetime type for ease of manipulation. In addition, columns storing information for the year, month, day as well as hour, minute and second are extracted from them, respectively, and the added to the original dataset.

Since the dataset contains a number of categorical variables, these are concerted into binary ones using the pandas get\_dummies() function. This process creates one column or feature for each level of each categorical variable and assigns 0s and 1s according to whether the level is attested in the record or not. This highly increases dimensionality in the dataset overall. For an alternative model using the H2O library, categorical features where used as such without turning them into binary features.

The last step of data wrangling involved resampling. Since the target variable is imbalanced in which one level accounts for 45% of all the data, both undersampling of the majority class and oversampling of the minority classes were attempted. This resampled data was used to run alternative models.

Create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.