Team MSDJ’s

Data Mining I – Spring 2017

Final Project

**INTRODUCTION**

**I. METHODOLOGY**

*A. Overview*

*B. Preprocessing*

In any application of data mining, the first thing that needs to be accomplished is processing the raw data into a workable form. The first step in accomplishing that goal, outside of changing the column names to less verbose tokens, was to convert seemingly numerical data into the factors described in the guide. The majority of the columns were factors, with only Number\_of\_Vehicles, Casualties, Lat/Long, and two derived columns remaining as non-factor data.

Next, the Time column as it existed was deemed unworkable. In its stead, a new categorical variable was implemented to declare the “Time Period”, a declaration of what 4-hour block of time a row corresponded to: Overnight (12-3:59am), Early Morning (4-7:59am), Morning (8-11:59am), Afternoon (12-3:59pm), Evening (4-7:59pm), and Night (8-11:59pm).

Following this, three more derived columns were created: isHoliday, isWeekend, and Month. Based on UK Bank Holidays from 2012, isHoliday (derived from Date) could lead to more frequent/higher severity accidents as people tend to drink in celebration. Month (also derived from Date) could potentially, when viewed hand-in-hand with Weather, highlight increased severity in the winter months. Similarly, isWeekend could show increased accident rates/severity as people may be more likely to drink on the weekends compared to work days.

These derived columns made it possible to drop Time and Date from the data, as well as a number of other intuitively useless columns: Index, E\_OSGR, N\_OSGR, LA\_Highway, Police\_Atttendance, and LSOA. Index, E\_OSGR, N\_OSGR, LA\_Highway, and LSOA were removed since they are location data for each observation, of which there was a glut. Further, Lat and Long were perfectly correlated with E\_OSGR and N\_OSGR. To lose one set over another was moot, so it was decided to keep the columns with more familiar terms. Even further, LSOA was a factor with 28987 levels. R was struggling to create such a high number of dummy variables, and a great deal of the LSOA entries were NA, so it was decided to remove the column as a whole in lieu of other location data. Finally, Police\_Attendance referred to whether or not an officer attended the scene of the accident after the fact. This was deemed to have little to no impact on predicting the outcome of an accident.

From there, the only columns describing NA values were Junciton\_Control, Second\_road\_class, and Surface\_Conditions. Removing such values from the data produced our first working data set, cleanData, which contained roughly 60% of the observations from our raw data set. However, some columns contained an “Unclassified” or “Unknown” level outside of the NA specified in the guide: First\_road\_class, Road\_Type, Second\_road\_class, Light\_Conditions, and Weather\_Conditions. These levels were reclassified in an even cleaner data set and removed, resulting in perfectData, containing only slightly over 17% of the observations from our raw data set.

It was recognized that losing 83% of the raw data was, perhaps, undesirable. As analysis of the data proceeded and insignificant predictors were identified, it was a goal to increase the number of workable observations while maintaining the integrity of the data.

*C. Summary of Method Performance*

*D. Final Model*

**II. KEY FINDINGS & MAIN TAKEAWAYS**

***A.*** *Figures*

***B****. Tables*

***C.*** *Output*

***D.*** *Discussion of Findings*

***E.*** *“Key Task” Question #2 – Based on the available data, what are the factors that best discriminate between different severities of accidents? Can the knowledge of these factors be helpful in the practice of reducing frequency of higher-severity accidents?*

***F.*** *“Key Task” Question #3 -* Can the given data be used to infer areas for improvement within different police jurisdictions? If so, provide the reasoning for your analysis with examples.