Assignment 5: Complete Analysis Using Watson Analytics

Daanish Ahmed
DATA 610 9080
Fall 2016
dsahmed2334@yahoo.com
Professor Vahe Heboyan

December 16, 2016

Introduction

In this course's previous assignments, I used IBM's Watson Analytics to perform several aspects of the decision-making process. I examined the data through an exploratory analysis, built predictive models through the use of decision trees, and finally communicated my findings by using displays. For this final assignment, my goal is to combine all three of these components to perform a complete analysis on a particular dataset. The dataset in question was found on the UK government's public database for road safety, and it contains accident information from the year 2015 ("Road Safety Data," 2011). This dataset is extremely robust since it has 32 input variables and an initial total of 140,056 cases (see Table 1 in Appendix A). This number was reduced to 76,950 cases after the removal of outliers, errors, and missing values. Some of the input variables include road type, weather conditions, junction detail, number of vehicles, and number of casualties. The variables that can be used to develop predictive models are accident severity and whether police attended the scene of the accident. The primary target for this assignment is accident severity, but I also intend to develop models for police attendance as well. The goal is to conduct a complete analysis to determine the factors that influence these two variables, and to communicate these findings with my organization. By understanding the causes of these accidents, I aim to help my potential organization to reduce the likelihood and severity of traffic collisions.

Data Exploration

First of all, I will explore the dataset by examining the initial questions provided by Watson Analytics. The first visualization is a spiral graph which shows the factors that most strongly influence the speed limit. By far, the strongest indicator is whether or not the accident site is in an urban or rural area—as speed limits are generally higher in rural areas. Further exploration reveals that rural areas also have more severe accidents (see Figure 3 in Appendix A). The second visualization is a line graph that shows average speed limits over the days of the week, all while considering whether an officer shows up at the accident site (see Figure 1 in Appendix A). The graph reveals that speed limit does not change significantly on different days of the week, but it more interestingly shows that officers are more likely to attend the site when accidents occur at higher speeds. The third visualization is an area curve that shows the average number of casualties per day of the week, also while considering officer attendance (see Figure 2 in Appendix A). The results indicate that slightly more accidents occur on weekends, and that officers are more likely to show up when there are higher numbers of casualties. Combined with the findings of the previous graph, these results suggest that officers are more likely to arrive when the accidents are more severe.

After exploring the initial questions, I created some specific questions to develop further insights (see Figure 3 in Appendix A). One of the visualizations is a stacked bar graph showing the breakdown of weather conditions based on accident severity. According to the results, foggy or misty conditions generally have the highest rates of serious or fatal accidents. Interestingly, some conditions such as "snowing with high winds" have surprisingly low rates of severe accidents. Another image shows a series of pie charts for accident severity with different road surface conditions. Here, we see that flooding leads to the highest number of severe accidents. But once again, conditions such as frost or snow lead to relatively low severity rates. This could

be due to the fact that some of these columns have smaller sample sizes compared to others. Other possibilities include fewer vehicles on the road, or drivers using more caution when driving under certain road or weather conditions. The last image shows a packed bubble that includes accidents for different junction details. I improved the relevancy by using filters to hide accidents of lower severity. The results indicate that T or staggered junctions are by far the most dangerous junction types. Overall, these answers can be valuable for using in a transportation, traffic, or road safety organization. By understanding some of the factors contributing to deadlier accidents, the organization will be able to take steps towards preventing them.

Data Refinement

In this section, I will describe my efforts to improve the quality of the data. In order to make the data more useful for analysis, I had to replace some of the entries in the dataset. Many of the variables used numeric values like 1 or 2 to represent the levels of accident severity or the road type. Using a provided key, I replaced these numbers with their corresponding text values ("Accident Statistics," 2011). After importing the dataset, I found that its data quality score was 61%. The variable of lowest quality—the accident date—had a score of 40% because some of its entries were of different date formats, causing Watson Analytics to think that those values were missing. By filtering out this column, the quality increased by 1%. The day of the week has the highest quality at 92% because its values are the most evenly distributed. Several other columns, such as carriageway hazards and weather type, have quality scores as low as 50% because the majority of their entries are of the same value. Though filtering out these columns will improve the data quality, it will also hide these columns from any discovery set in which they are used. For the sake of the analysis, I chose not to hide these columns. It is also possible to create hierarchies in this dataset—but since most of these variables are categorical and do not overlap with each other, any created hierarchy would make little sense and have limited analytical value. Data groups, on the other hand, are much more useful here. For example, I created a grouping called "light condition type," simplifying the values for lighting conditions into either "light" or "dark." These groupings allow me to filter out some columns with low data quality and improve the overall quality score.

Decision Trees

In this part of the paper, I will discuss the predictive models that I created to determine accident severity and police attendance of accident sites. Firstly, the decision tree on accident severity is mainly meant to predict serious accidents. To simplify this decision tree, I clipped the branches with the lowest number of severe or fatal accidents, including all cases where police did not show up at the accident site (see Figure 4 in Appendix A). Doing so resulted in a tree that is both easier to understand and more specific since it hides some of the less severe cases. Some of the strongest predictors include police presence at the accident, size of the police force, and number of vehicles involved. The single best predictor is police attendance, since their presence often indicates a more severe accident. The best combination of variables is the size of the police force and their presence at the accident. Larger police forces in addition to attendance at the accident both correspond with more dangerous accidents. Using the top three decision rules

(see Figure 5 in Appendix A), I noticed that these cases mostly occurred on dark streets, and that police officers were almost always present. In addition, most of the serious accidents occur on single carriageway roads, which have been considered to be seven times more dangerous than larger highways (Millward, 2013). If I worked for a transportation or traffic organization, analyzing these rules would be a priority. These rules show the most likely cases for severe accidents occurring, and thus my organization should focus on either avoiding certain road conditions or improving its transportation or roads system based on these findings.

Afterwards I created a decision tree on officer attendance, this time focusing on cases where an officer is least likely to show up. I simplified the model by clipping branches with high police attendance rates, including cases with high casualties (see Figure 6 in Appendix A). The resulting decision tree has fewer variables and is easier to understand. The focus on no police attendance means that the tree generally focuses on less severe accidents, thus making it different from the previous model. Some of the strongest predictors include the number of vehicles, speed limit, police force size, and junction detail. The single best predictor is the number of vehicles, indicating that police are less likely to attend when more vehicles are present in an accident. Further exploration reveals that accidents with higher numbers of vehicles tend to be less lethal. The best combination of variables is that of accident severity and police force size—cops are less likely to arrive at the scene if the police force is small and the accident is less severe. By looking at the top three decision rules (see Figure 7 in Appendix A), I observed that these cases always involve accidents of low severity where casualties were minimal or nonexistent. Likewise, speed limits in these cases are relatively low while road conditions are usually dry. My organization should study these rules to learn how to lower accident rates and thus allow police forces to focus attention to where they are needed the most.

Displays

In the final paragraphs of this paper, I will describe the displays that I created in order to communicate the results of my analysis. The first display is a dashboard on accident severity that includes visualizations from my exploratory analysis (see Figure 3 in Appendix A). This dashboard is designed to tell the story of how different weather conditions or road types can affect the severity of accidents. The first image shows that accidents are more likely to be dangerous in rural areas rather than urban ones. The second and third images describe weather and road conditions, and the findings suggest that flooding and foggy weather both result in the highest number of severe accidents. As mentioned earlier, these images also reveal that snowy or icy conditions contribute to the lowest numbers of serious accidents—possibly due to either higher levels of driver caution, smaller data sample sizes, or fewer vehicles on the road. The final image suggests that certain junction types—such as T or staggered junctions—are more hazardous than others. One possible application for my organization could be the design of safer roads. Having wider and better-maintained roads and junctions—or even lower speed limits—could significantly reduce accident severity in rural areas or under certain weather conditions. Some people may argue that my display has a narrow focus and does not consider other possibilities, but I would advocate its use as a starting point for improving road safety.

The second dashboard that I created focuses on police presence at the accident scene (see Figure 8 in Appendix A). The first image shows the likelihood of police arriving based on the

accident severity level. It reveals that police are almost guaranteed to arrive at fatal accidents, but there is only an 80% chance that they will arrive at minor accidents. The second visualization focuses on police attendance based on the junction control type (which includes traffic lights and stop signs). One finding is that police are least likely to arrive at accidents occurring at stop signs. A possible explanation is that accidents at stop signs often occur at lower speeds compared to traffic lights or other junction types. The third image shows the number of accidents occurring at certain daylight conditions, based on police attendance. The graph reveals that most accidents occur during the daytime, but by using filters I noticed a slight increase in police attendance during nighttime accidents. The last visualization consists of police presence for serious and fatal accidents based on the number of casualties. Here, we see that police presence is almost 100% likely for accidents with higher numbers of casualties. If my organization wants to focus on safety, then this display offers a good approach on how to promote safe driving. The combination of the two displays developed so far provides some helpful tips on lowering the likelihood of road collisions.

I then developed one final dashboard display, this time using data imported from Twitter (see Figure 9 in Appendix A). My goal was to relate Twitter data to my findings on accident severity and weather conditions. Some of the hashtags that I included are accident, police, flood, fog, ice, rain, snow, and wind. The image on the upper right is a map which reveals that the countries with the highest number of relevant tweets are the U.S., Canada, and the U.K. The image on the lower right serves as a timeline of tweets per day that include the matching hashtags. This graph reveals that snow is tweeted more often than any other hashtag combined. The visualization on the left provides a breakdown of the most common hashtags for the top three countries. This image reveals that snow by far the most common hashtag appearing in the U.S. and Canada. For the U.K., however, there are a large number of tweets about the police—yet there are few tweets regarding road accidents. I used filters to include only tweets related to accidents, and I found very few tweets that linked accidents to either weather or police presence. As a result, my findings were not as relevant to the subject matter as I would have hoped. Nevertheless, I believe that this method can still be useful for my organization because my efforts in previous assignments have yielded more insightful results. However, users need to be careful and select hashtags that are more commonly used by Twitter users.

Finally, I organized the three dashboards that I created so far into a storybook display (see Figures 10, 11, and 12 in Appendix A). I included comments and annotations for each image in order to help viewers understand the visualizations in my analysis. Here, my goal is to provide a complete overview of my analysis to help my organization understand how it can improve transportation safety. This display format is helpful because it groups all three of my previous displays together in a coherent manner. The structure of this display allows users to interact with my visualizations and formulate key insights from each graph. Additionally, it is possible for viewers to replace my dataset with their own; though I would recommend using a dataset of a similar format to the one I used because doing otherwise may generate errors due to mismatched columns. One possible shortcoming from using my approach is that the storybook focuses only on a few key takeaways and does not provide the complete picture to understanding road accidents. Thus, some in my organization may resist the adoption of this approach. However, I would argue that a complete understanding would only happen if we rely more heavily on a data-driven approach. By expanding our knowledge of the data, my organization will be better equipped towards making decisions that could potentially save lives on the road.

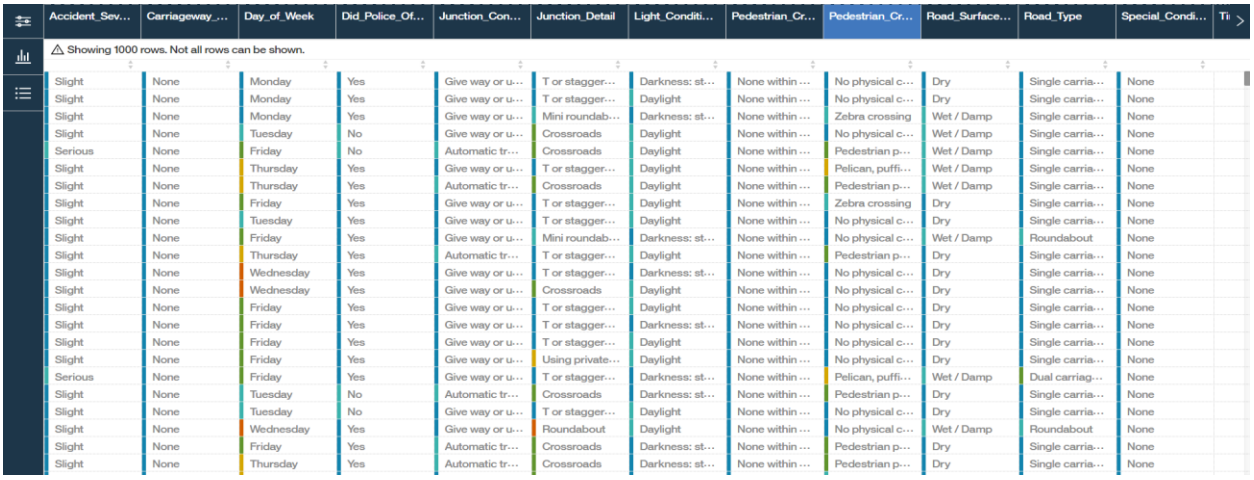
References

Accident Statistics. (2011, September). Retrieved December 8, 2016, from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230590/stats19.pdf

Millward, D. (2013, October 24). Single carriageway roads seven times more dangerous than motorways. *The Telegraph.* Retrieved December 12, 2016, from http://www.telegraph.co.uk/news/uknews/road-and-rail-transport/10400682/Single-carriageway-roads-seven-times-more-dangerous-than-motorways.html

Road Safety Data. (2011, September 20). Retrieved December 7, 2016, from https://data.gov.uk/dataset/road-accidents-safety-data

Appendix A



Accident_Sev...	Carriageway_...	Day_of_Week	Did_Police_Of...	Junction_Con...	Junction_Detail	Light_Conditi...	Pedestrian_Cr...	Pedestrian_Cr...	Road_Surface...	Road_Type	Special_Condi...	Ti
⚠ Showing 1000 rows. Not all rows can be shown.												
Slight	None	Monday	Yes	Give way or u...	T or stagger...	Darkness: st...	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Monday	Yes	Give way or u...	T or stagger...	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Monday	Yes	Give way or u...	Mini roundab...	Darkness: st...	None within ...	Zebra crossing	Wet / Damp	Single carria...	None	
Slight	None	Tuesday	No	Give way or u...	Crossroads	Daylight	None within ...	No physical c...	Wet / Damp	Single carria...	None	
Serious	None	Friday	No	Automatic tr...	Crossroads	Daylight	None within ...	Pedestrian p...	Wet / Damp	Single carria...	None	
Slight	None	Thursday	Yes	Give way or u...	T or stagger...	Daylight	None within ...	Pelican, puffi...	Wet / Damp	Single carria...	None	
Slight	None	Thursday	Yes	Automatic tr...	Crossroads	Daylight	None within ...	Pedestrian p...	Wet / Damp	Single carria...	None	
Slight	None	Friday	Yes	Give way or u...	T or stagger...	Daylight	None within ...	Zebra crossing	Dry	Single carria...	None	
Slight	None	Tuesday	Yes	Give way or u...	T or stagger...	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Friday	Yes	Give way or u...	Mini roundab...	Darkness: st...	None within ...	No physical c...	Wet / Damp	Roundabout	None	
Slight	None	Thursday	Yes	Automatic tr...	T or stagger...	Daylight	None within ...	Pedestrian p...	Dry	Single carria...	None	
Slight	None	Wednesday	Yes	Give way or u...	T or stagger...	Darkness: st...	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Wednesday	Yes	Give way or u...	Crossroads	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Friday	Yes	Give way or u...	T or stagger...	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Friday	Yes	Give way or u...	T or stagger...	Darkness: st...	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Friday	Yes	Give way or u...	T or stagger...	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Friday	Yes	Give way or u...	Using private...	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Serious	None	Friday	Yes	Give way or u...	T or stagger...	Darkness: st...	None within ...	Pelican, puffi...	Wet / Damp	Dual carriag...	None	
Slight	None	Tuesday	No	Automatic tr...	Crossroads	Darkness: st...	None within ...	Pedestrian p...	Dry	Single carria...	None	
Slight	None	Tuesday	No	Give way or u...	T or stagger...	Daylight	None within ...	No physical c...	Dry	Single carria...	None	
Slight	None	Wednesday	Yes	Give way or u...	Roundabout	Daylight	None within ...	No physical c...	Wet / Damp	Roundabout	None	
Slight	None	Friday	Yes	Automatic tr...	Crossroads	Darkness: st...	None within ...	Pedestrian p...	Dry	Single carria...	None	
Slight	None	Thursday	Yes	Automatic tr...	Crossroads	Darkness: st...	None within ...	Pedestrian p...	Dry	Single carria...	None	

Table 1. 2015 UK Accidents Dataset.

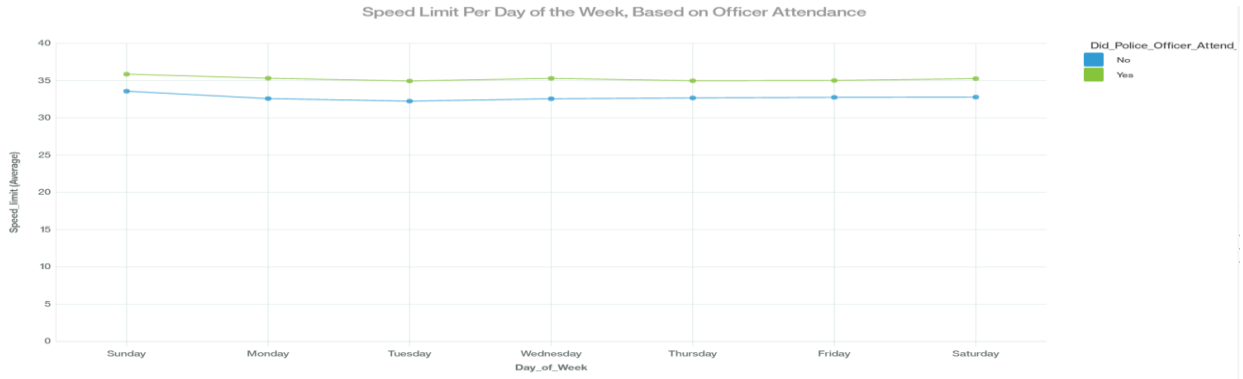


Figure 1. Average Speed Limit Per Days of the Week Based on Officer Attendance.

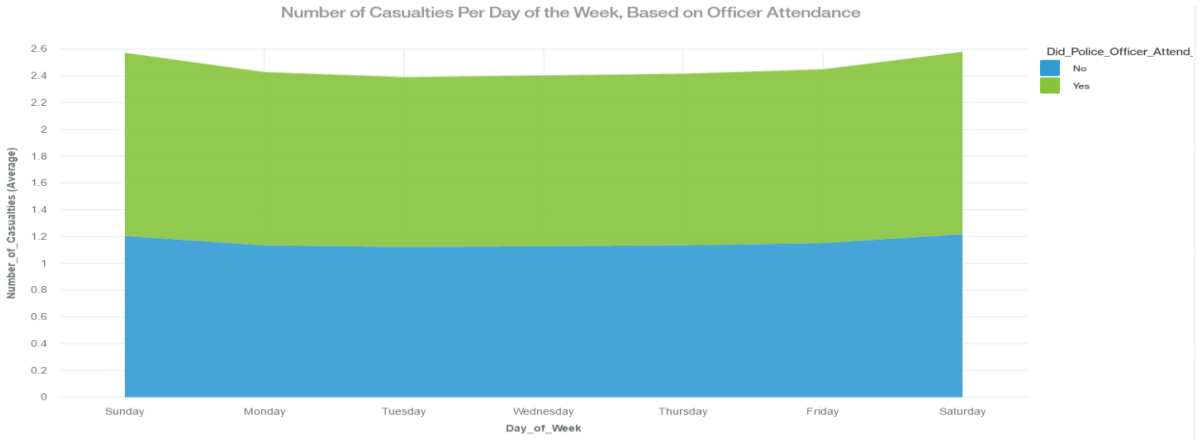


Figure 2. Average Number of Casualties Per Day Based on Officer Attendance.

Figure 3. Accident Severity Dashboard and Exploratory Questions.

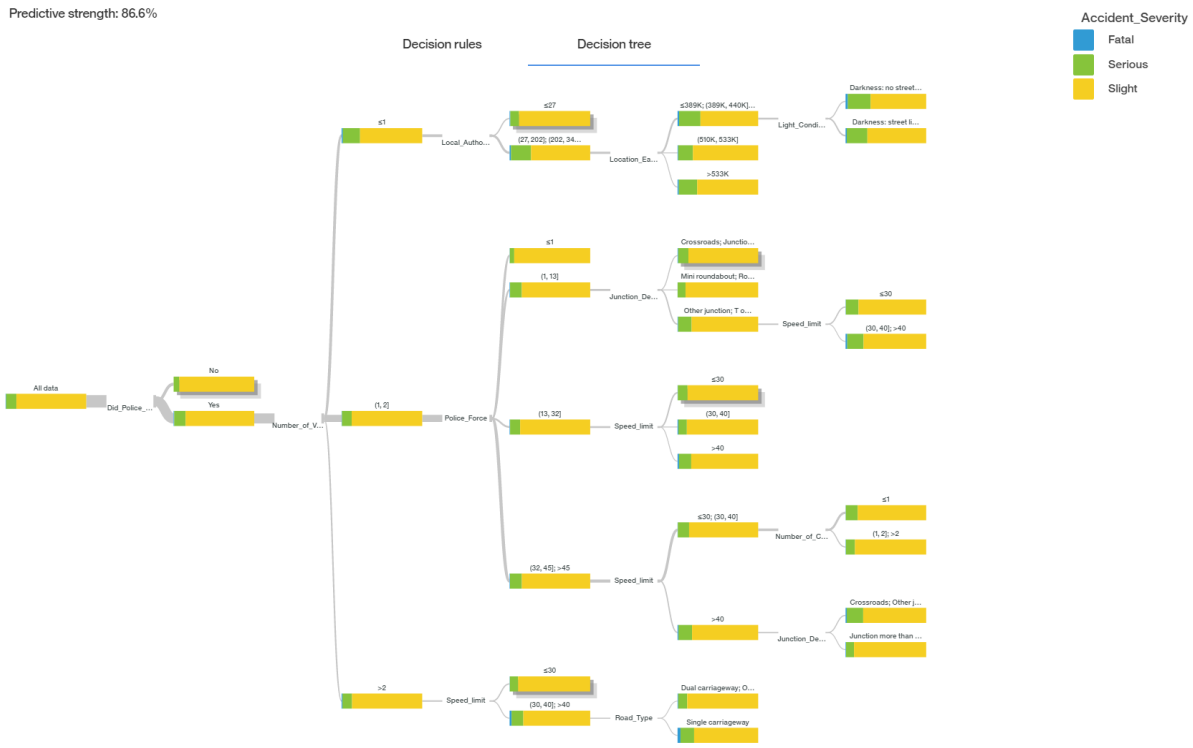


Figure 4. Accident Severity Decision Tree.

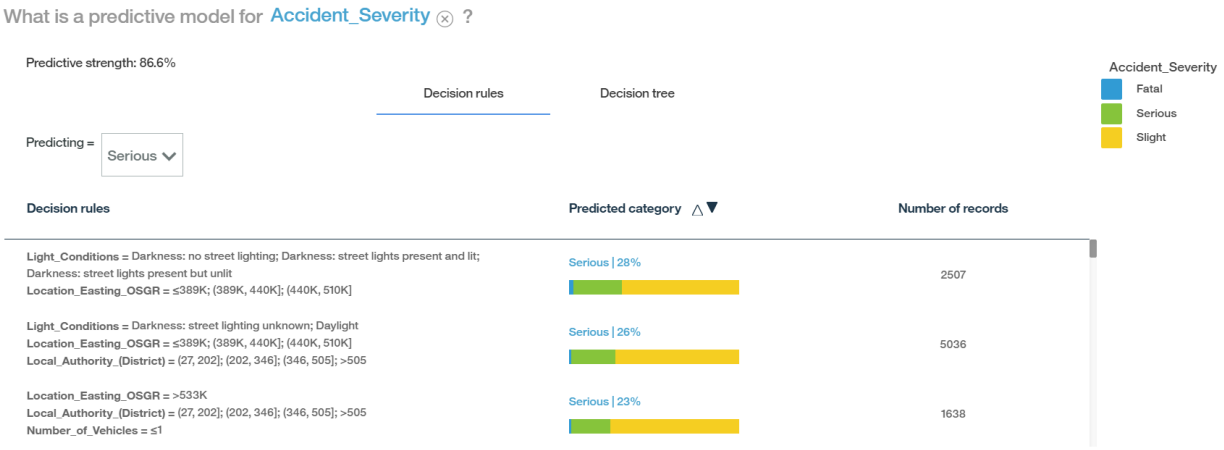


Figure 5. Top Decision Rules for Serious Accidents.

Figure 6. Police Attendance at Accident Site Decision Tree.

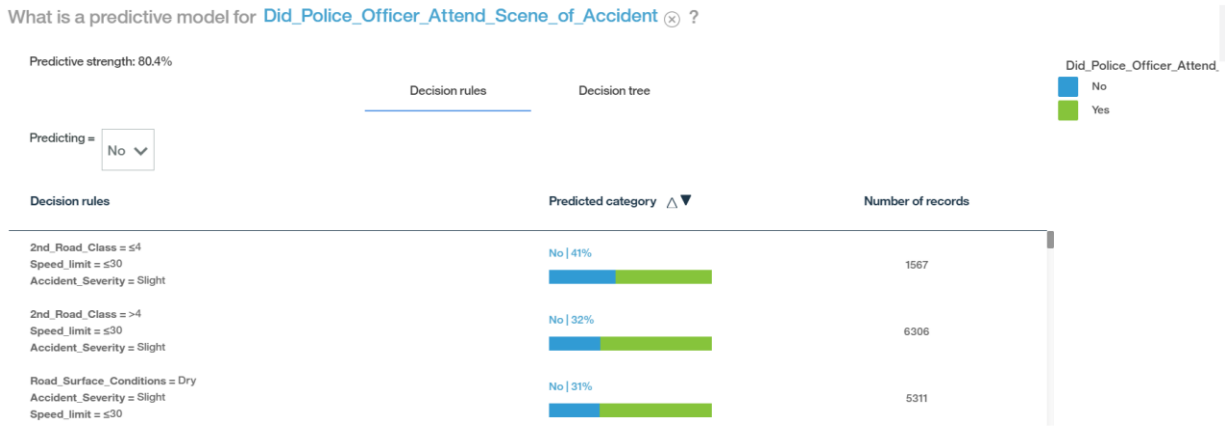


Figure 7. Top Decision Rules for Police Not Attending Accident Site.

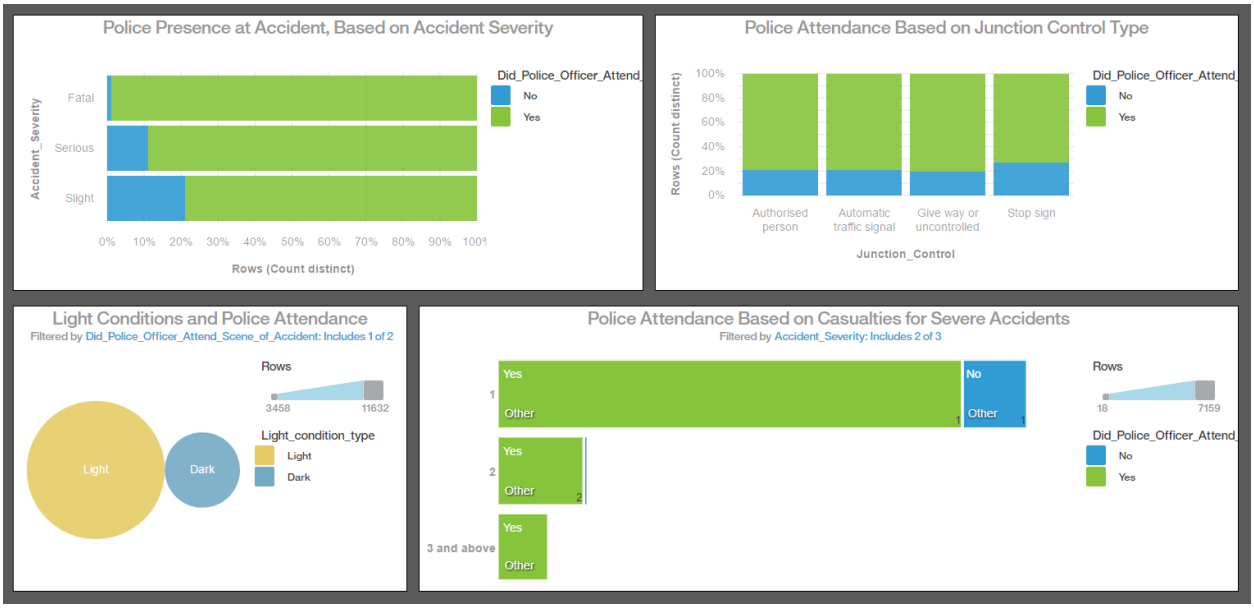


Figure 8. Police Attendance Dashboard.

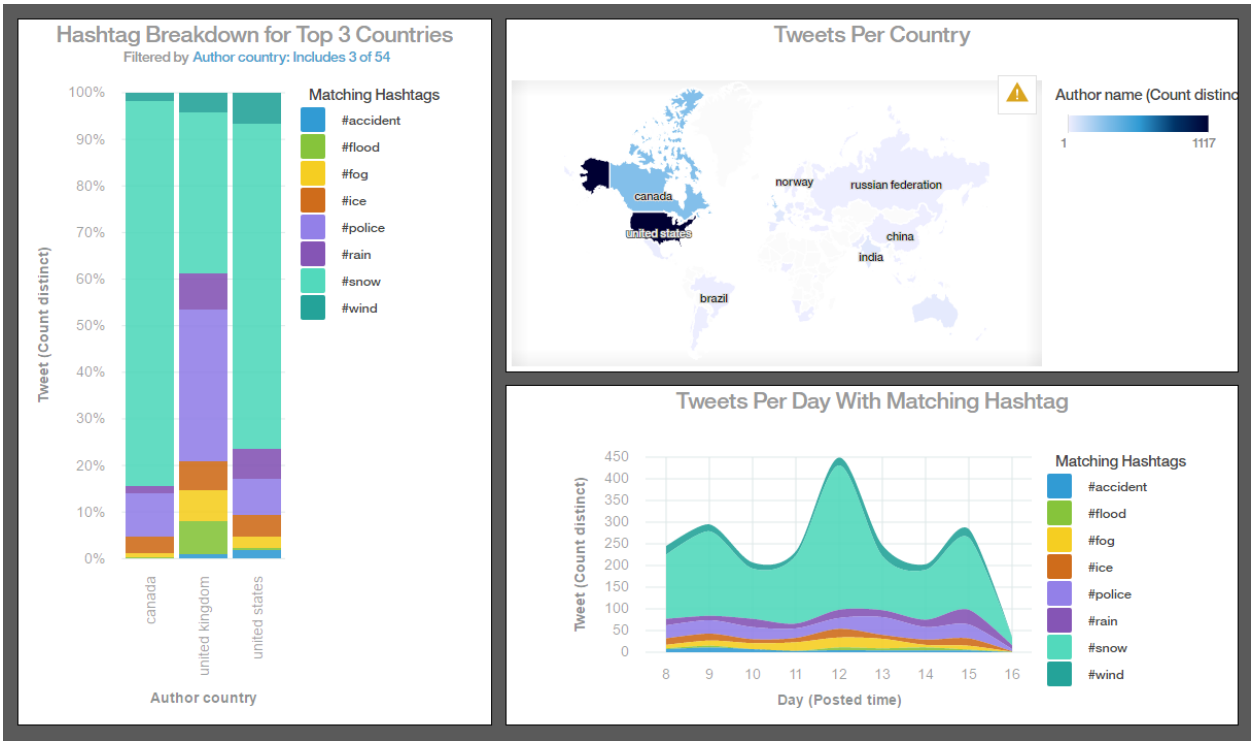


Figure 9. Twitter Dashboard for Accident and Road Conditions.

Figure 10. Accident Severity Storybook Page.

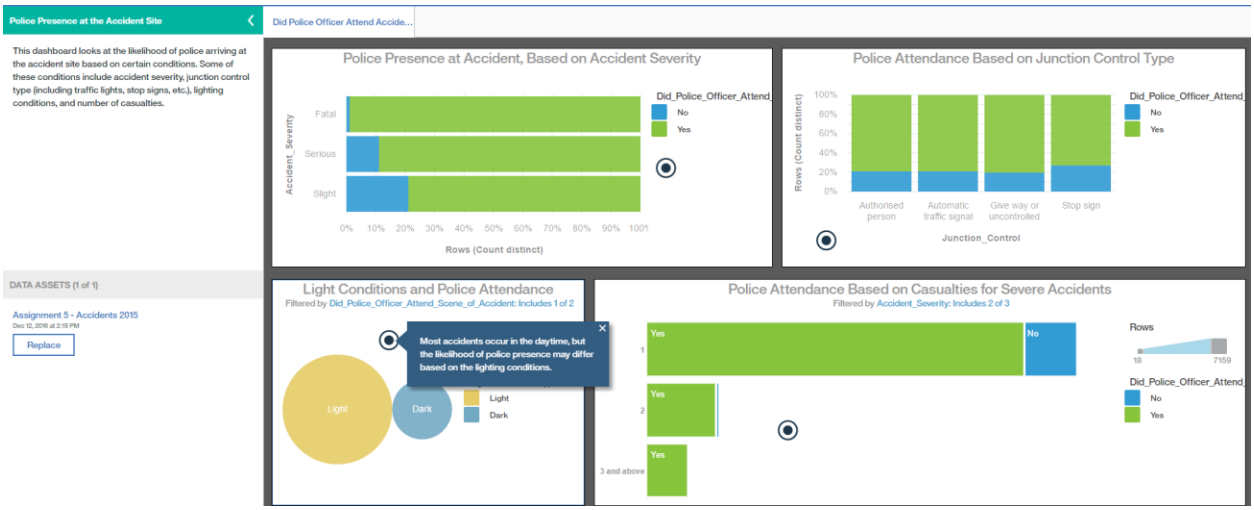


Figure 11. Police Attendance at Accident Site Storybook Page.

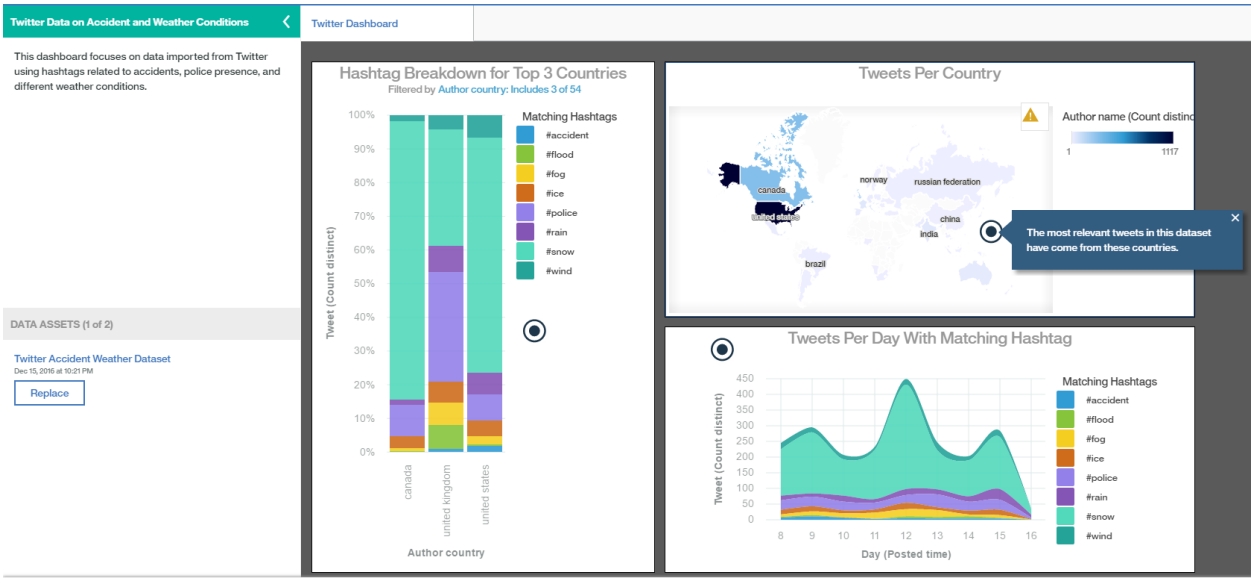


Figure 12. Twitter Storybook Page.