Madhukar Mantravadi
Data Science Position Research Test
1/31/2019

QUESTION 1: Causal Hypothesis

Looking at the data the more specific/detailed features such as location and street name have far more categories than some of the other more general ones such as whether the driver is at fault for the accident or not. Because of the amount of unique values, I decided to look into the effect of time of day via the 'Light' feature on Injury Severity and Vehicle Damage Extent

Hypothesis: The light will affect the Injury Severity and Vehicle Damage Extent and that the darker it is, the higher the ratio of severe damage to temperate damage.

First, we must look at just the former part of the hypothesis. Does the light have a relationship with damage at all?

Null Hypothesis (Ho) = The light (time of day) doesn't have a relationship with the Injury Severity and Vehicle Damage Extent

Alternative hypothesis ($H1$) = The light (time of day) does have a relationship with the Injury Severity and Vehicle Damage Extent

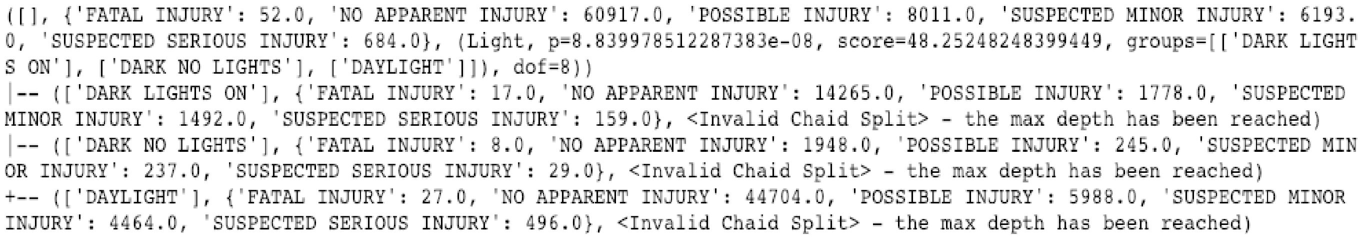
QUESTION 2: Causal Inferential Model

1. To do this experiment, first $\chi 2$ test will be done to determine which hypothesis to reject. This will find the existence of a relationship if there is one.

2. The categories in the data that had much smaller value counts were removed to simplify the problem.

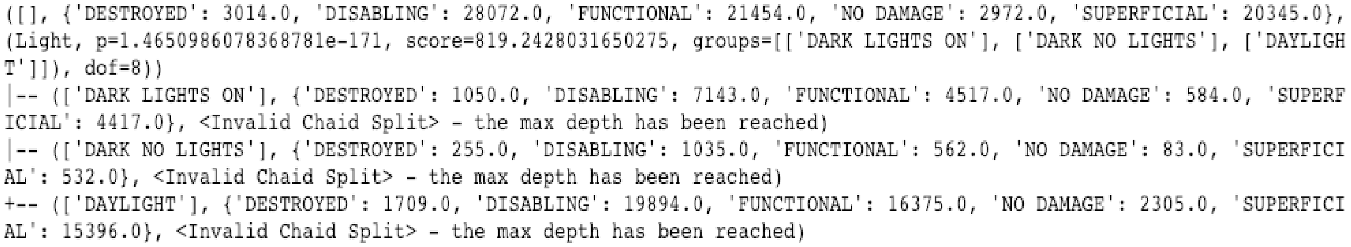
3. Masks were created to filter the data based on the values that were to be cut and then also dropped observations that were not defined in the data, resulting in NaN values.
4. The CHAID (Chi-squared Automatic Interaction Detector) algorithm was used to do the test. An existing library called CHAID was used to generate the values easily.
    a. The CHAID Algorithm checks the statistical significance via the $\chi 2$ test and outputs the counts of all categories within each of the features of the variable. It does this for every variable and generates every crosstab it can that has statistical significance.
    b. If a feature is shown to not be significant, it will try to combine it with another non-significant feature and run the test again. This can lead to misleading or useless results for certain features. In the analysis steps were taken to avoid this, but it is something to be weary of. One should not just follow a model's results blindly.

CHAID Tree printed out for the Injury Severity feature



```
([], {'FATAL INJURY': 52.0, 'NO APPARENT INJURY': 60917.0, 'POSSIBLE INJURY': 8011.0, 'SUSPECTED MINOR INJURY': 6193.
0, 'SUSPECTED SERIOUS INJURY': 684.0}, (Light, p=8.839978512287383e-08, score=48.25248248399449, groups=[['DARK LIGHT
S ON'], ['DARK NO LIGHTS'], ['DAYLIGHT']]), dof=8))
|-- (['DARK LIGHTS ON'], {'FATAL INJURY': 17.0, 'NO APPARENT INJURY': 14265.0, 'POSSIBLE INJURY': 1778.0, 'SUSPECTED
MINOR INJURY': 1492.0, 'SUSPECTED SERIOUS INJURY': 159.0}, <Invalid Chaid Split> - the max depth has been reached)
|-- (['DARK NO LIGHTS'], {'FATAL INJURY': 8.0, 'NO APPARENT INJURY': 1948.0, 'POSSIBLE INJURY': 245.0, 'SUSPECTED MIN
OR INJURY': 237.0, 'SUSPECTED SERIOUS INJURY': 29.0}, <Invalid Chaid Split> - the max depth has been reached)
+-- (['DAYLIGHT'], {'FATAL INJURY': 27.0, 'NO APPARENT INJURY': 44704.0, 'POSSIBLE INJURY': 5988.0, 'SUSPECTED MINOR
INJURY': 4464.0, 'SUSPECTED SERIOUS INJURY': 496.0}, <Invalid Chaid Split> - the max depth has been reached)
```

CHAID Tree printed out for the Vehicle Damage Extent feature

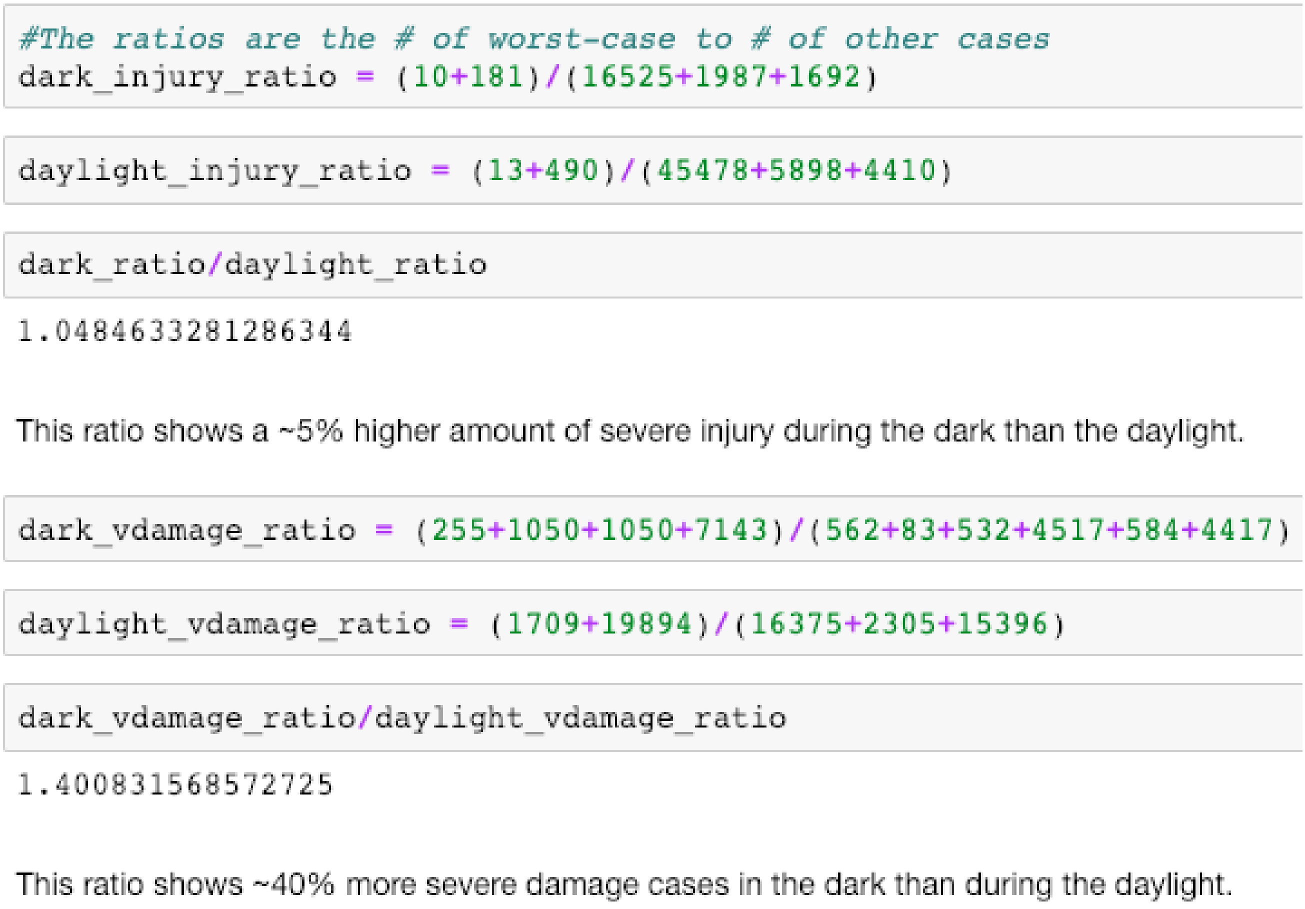


```
([], {'DESTROYED': 3014.0, 'DISABLING': 28072.0, 'FUNCTIONAL': 21454.0, 'NO DAMAGE': 2972.0, 'SUPERFICIAL': 20345.0},
(Light, p=1.4650986078368781e-171, score=819.2428031650275, groups=[['DARK LIGHTS ON'], ['DARK NO LIGHTS'], ['DAYLIGH
T']]), dof=8))
|-- (['DARK LIGHTS ON'], {'DESTROYED': 1050.0, 'DISABLING': 7143.0, 'FUNCTIONAL': 4517.0, 'NO DAMAGE': 584.0, 'SUPERF
ICIAL': 4417.0}, <Invalid Chaid Split> - the max depth has been reached)
|-- (['DARK NO LIGHTS'], {'DESTROYED': 255.0, 'DISABLING': 1035.0, 'FUNCTIONAL': 562.0, 'NO DAMAGE': 83.0, 'SUPERFICI
AL': 532.0}, <Invalid Chaid Split> - the max depth has been reached)
+-- (['DAYLIGHT'], {'DESTROYED': 1709.0, 'DISABLING': 19894.0, 'FUNCTIONAL': 16375.0, 'NO DAMAGE': 2305.0, 'SUPERFICI
AL': 15396.0}, <Invalid Chaid Split> - the max depth has been reached)
```

5. The **p-value** for *Injury Severity* =8.8399**e**−08=8.8399e−08
   The **p-value** for *Vehicle Damage Extent* =1.4651**e**−171=1.4651e−171
   Both of these values are <0.05∴ statistically significant.

6. As we can see above both of the cases have a **p value < 0.05** and are therefore statistically significant. We can also see this by the fact that the $\chi 2$ score is relatively far from 0 (especially in the case of vehicle damage).
7. We are part way into proving our hypothesis true or false but within the context of our $\chi 2$ test, we can reject the null hypothesis and accept the alternate hypothesis that the Light (time of day) has a relationship with Vehicle Damage Extent and Injury Severity.
8. Below I have calculated the ratio of worse categories to more temperate ones. I assumed that for Injury that "Fatal Injury" and "Suspected Serious Injury" were the worst-case categories. For Vehicle Damage Extent, I assumed "Destroyed" and "Disabling" were the worst-case. The ratio's below are a simple calculation showing that our original hypothesis has grounds to be correct based on the data. However, correlation is not causation, so we cannot assume that this proves the latter part of our hypothesis that more serious accidents happen in the dark.

Calculations



```
#The ratios are the # of worst-case to # of other cases
dark_injury_ratio = (10+181)/(16525+1987+1692)
```

```
daylight_injury_ratio = (13+490)/(45478+5898+4410)
```

```
dark_ratio/daylight_ratio
```

```
1.0484633281286344
```

This ratio shows a ~5% higher amount of severe injury during the dark than the daylight.

```
dark_vdamage_ratio = (255+1050+1050+7143)/(562+83+532+4517+584+4417)
```

```
daylight_vdamage_ratio = (1709+19894)/(16375+2305+15396)
```

```
dark_vdamage_ratio/daylight_vdamage_ratio
```

```
1.400831568572725
```

This ratio shows ~40% more severe damage cases in the dark than during the daylight.

Question 3: Predictive Hypothesis

An interesting feature was the 2nd Harmful Event feature. This feature inspired an idea for a predictive model that could predict the existence of a 2nd Harmful Event based on Time of Day (Hr) and Place (Lat/Long). The idea for this was that if authorities were able to predict the existence of a 2nd Harmful event they could potentially allow law enforcement the benefit of knowing which accidents sites were most likely to be more serious and preemptively undertake preparations for a worse incident in terms of damage.

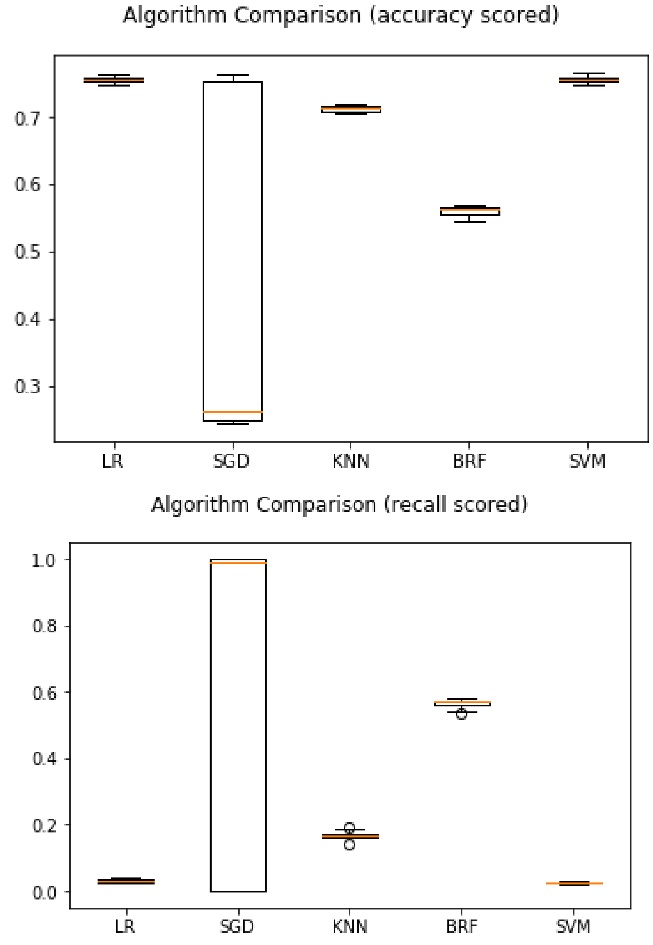
The reasoning on the features chosen was that time was a seen as a heavily influencing factor in the frequency of accidents occurring (see figure below), and location was chosen because generally speaking, more heavily trafficked areas are bound to result in higher frequencies of incidents. The type of First Harmful Event was also chosen because by looking at the types and also just using intuition, one can imagine that certain types of incidents have higher probability of causing a 2nd Harmful Event. For example, one would assume that a simple rear end may not have the same probability for a 2nd Event as say an incident with a pedestrian.

The hypothesis in this case is that the Time, Place, and the type of Harmful Event that came first are enough to predict the existence of a 2nd Event confidently.

Question 4:

1. For this problem I chose multiple models to see which method will work better. The specific models were chosen because they can deal with unbalanced data and also have penalties in place, so they can try to avoid falling to either end of the bias-variance spectrum.
2. The data was split into a training and test set first, so that a holdout set would be ready to test the final model's strength.
3. I wrote a function that tested out multiple algorithms and output the different scores and standard deviations for accuracy and recall (because in the scenario of law enforcement, false negatives matter more heavily). The models I chose were: Logistic Regression, SGD (Stochastic Gradient Descent) Classifier, K-Nearest Neighbors, Balanced Random Forest (a flavor of RF that has an inbuilt sampling and bootstrapping method to help with unbalanced data), and SVC (Support Vector Classifier). These algorithms were chosen because they do better with unbalanced classification than other models. I did not want to go too complex either because interpretability and computational complexity is important to keep in mind.

Algorithm Comparison Boxplots

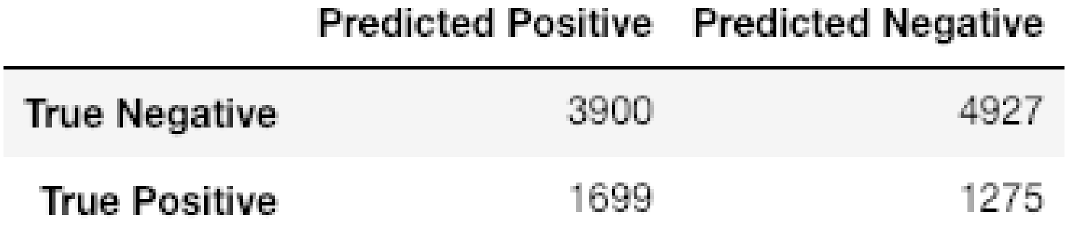


Algorithm Comparison (accuracy scored)

Algorithm Comparison (recall scored)

4. The Plots above show the resulting scores in a boxplot format. Firstly, SGD's standard deviation was very high, a telltale sign of overfitting and too high variance. SGD is an algorithm that is known to be very sensitive, but it was used because of a good track record as a classifier for unbalanced data. Of all the models, The Random Forest model performed reasonably when scored for both accuracy and recall. Again, recall is important because in the context of accidents, predicting an accident is less serious (there is less damage/no 2nd event) when it actually is can be very bad.
5. The Random Forest model had further hyperparameter tuning utilizing GridsearchCV from the sklearn python library.

Confusion Matrix of final model



	Predicted Positive	Predicted Negative
True Negative	3900	4927
True Positive	1699	1275

6. The confusion matrix above proves our hypothesis fairly wrong. Although the model does predict the possibility of a 2nd harmful event occurring, it does so with considerably low accuracy. The model has a large number of false positives because recall was chosen as the metric to focus on. But even so, there are still a lot of false negatives, which makes the model not useful. The point of the model is to try and predict when authorities need to employ extra resources to a crash because of the existence of a secondary harmful event. This model would not perform well in the real world and needs more features in training to be useful.

Question 5:

I believe that for the causal hypothesis of the relationship between light and severity of crashes there isn’t much more that is needed past what was given necessarily. More information about the specifics of the locations would be helpful, i.e. the size of the roads (wide/narrow), if there was a shoulder, if there was a sidewalk, what the area was like (wooded, rural, residential/surrounded by buildings). The data might be captured in some form of real-estate document or government documents about roads in the county (probably a construction database). The existence of these extra data would need to be researched because some may be harder to find than others.

The size of the road would help this analysis in being able to add another dimension in predicting accidents. It could help the model realize that maybe narrower roads have more likelihood to have an accident, or perhaps it’s less, only the data can tell.

The shoulder data would be a good baseline safety assumption especially because one of the types of incidents has to do with dividers/barriers.

The sidewalk data would help in cases with pedestrians and determining pedestrian traffic along with normal traffic and how it effects the likelihood of a crash and the severity of that crash. Will someone be more likely to swerve and cause a larger accident because there was someone walking on the sidewalk who steps into the street? These are questions we might be able to answer to some degree in a statistically sound way.

Question 6:

1. The Data:
    a. The Data would consist of all the minor traffic infractions/crimes committed
    b. Examples of these crimes would be things like speeding, running stoplights, etc.
2. Collection:
    a. The collection would most likely be via an excel dump from the police dept (or SQL Server depending on what system they use). If it is an SQL system, an SSIS package job could be scheduled.
    b. Assuming they are using a database of some kind, scheduling an import of data from a standard directory they have set up on their end would not be difficult as scheduling is a feature in several databases.
    c. This method would take some software to do so i.e. in SQL's case you need the SQL Server Management Studio
3. Frequency:
    a. Ideally it would be end of the day, however there could be a sweet spot of time that would suit better. For example, in the case of traffic incidents/car crashes the fewest number of cars on the road is at, say, 3 AM. Then the scheduling would be set for 3:30AM to ensure that the data is up to date and there aren't any entries that slip through until the next cycle.
4. Storage:
    a. As mentioned before it would most likely be stored in a database. Most likely it would be some form of relational database (Microsoft SQL, Oracle, RDS, etc.) There is also always the possibility of a NoSQL database (MongoDB, Redis, Neo4j) to be used for potentially easier maintenance and scalability. However, depending on the size of the data it may also be smart to invest in building out big data infrastructure, utilizing the Hadoop ecosystem (Spark, HQL/Hive, Pig, Ambari, etc.)
    b. AWS, Azure, GCP can also be leveraged in regard to big data. And in fact, if AWS is used, a data pipeline built entirely on their platform is possible, utilizing redshift/aurora/dynamo/s3 and other technologies within the AWS ecosystem.
    c. Ultimately it would depend on the form and volume of the output from the source, which especially in the case of a law enforcement organization, is not in one's control.

5. Analysis:
    a. The analysis would consist mostly of looking at the frequency of the types of infractions, the potential correlations between what infractions are most common and the places they happen in, and if there are any common denominators to the infractions either in an area, a certain period of time, or just in terms of do any of the infractions often “pair” together (i.e. running a stoplight and speeding)
    b. The analysis itself could be carried out in different ways depending on the system the data is stored on. It could be done via SSIS as mentioned before of scheduling a job that conducted the analysis, or perhaps in real-time utilizing streaming techniques such as Kafka (Apache) or Lambda (AWS).
6. Result:
    a. The findings would be reported most likely by a data visualization system such as Tableau/looker to schedule a dashboard to be output with the specified metrics and statistics. These among other software have a fairly comprehensive system to setup an automated/scheduled output such that the dashboard is updated on schedule and the report contains the proper data visualizations that are relevant to the current interests. And because Interests can change it isn’t too difficult to redo the scheduling with different visualizations, metrics etc.

Question 7:
I created some visualizations and put them in an html file in this folder. It should open up in any browser and display properly.

EXTRA:
I have also included the jupyter notebook file I was working in for most of the project if you want to see it.