Mini Project 1: AVs Safety Analysis

CS 498: Data Science & Analytics (Spring 2019)

Task 0

1. Imported data into Jupyter Notebook
   - Converted month to datetime

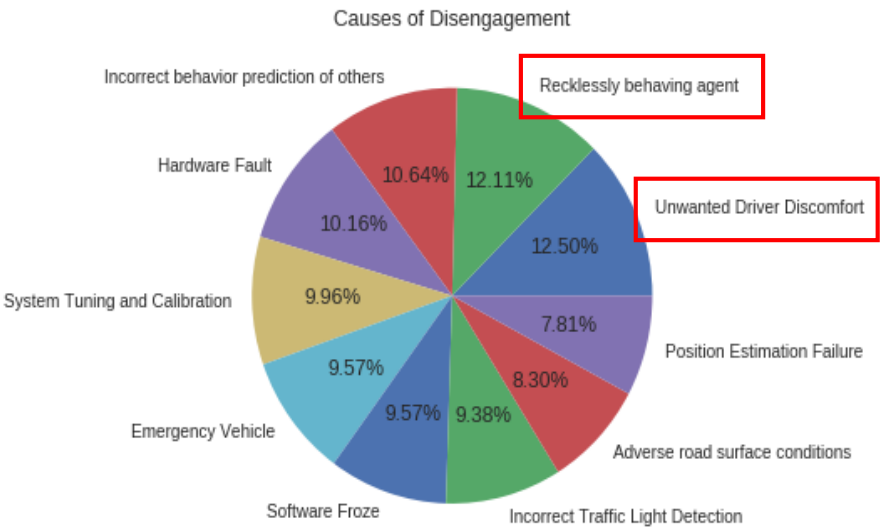
2. Summary

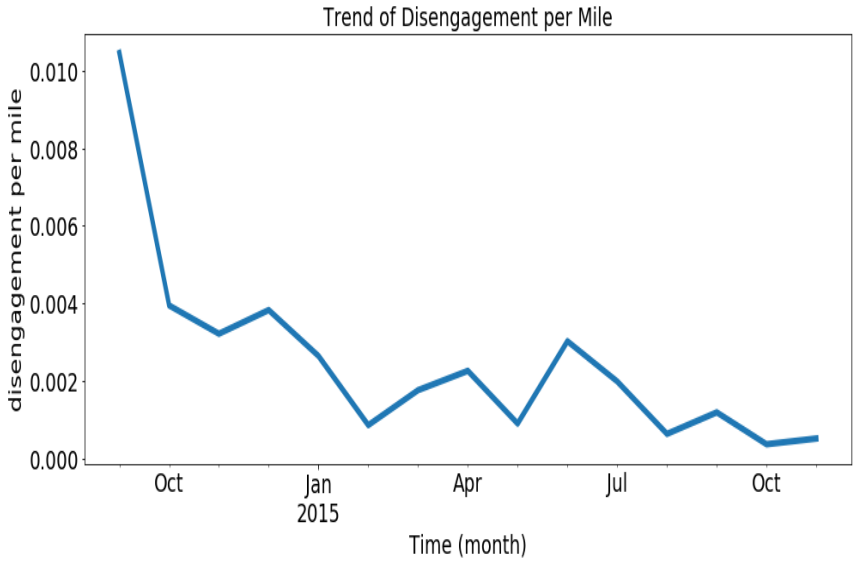
# of AV disengagements	1024
# of unique months	15
Unique Locations	urban-street, highway
# of unique causes	10
# of missing values	532

3. Pie Chart

   Top 2 Causes
   - Unwanted Driver Discomfort : 12.5%
   - Recklessly behaving agent: 12.11%

4. Trend of disengagement/mile
   - disengagement/mile is decreasing with time
   - Yes, AVs are maturing over time



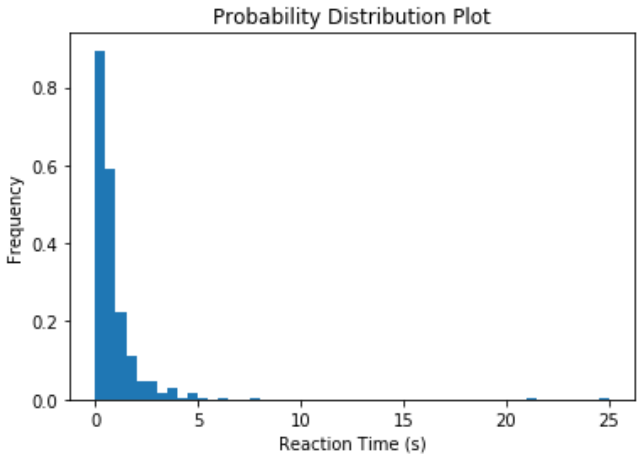


Task 1

1. Gaussian Distribution : $f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Exponential Distribution : $f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$

   Weibull Distribution : $f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$

2. The probability distribution of the reaction times fits the expo Weibull distribution.
   It signifies that the number of people with higher reaction time keeps decreasing as time increases

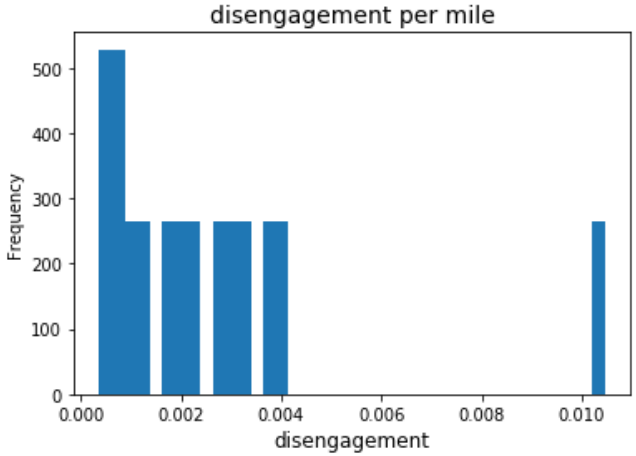


3. Average reaction time: 0.92977 seconds

Location	ReactionTime
highway	1.48000
urban-street	0.92865

4. Tested the hypothesis if the reaction time of humans in AV is different from non-AV at 0.05 significance level and concluded that yes, it is different in AV from non AV

5. The probability distribution of disengagement per mile fits the exponential distribution.
   It signifies that the probability of disengagement keeps decreasing as the distance increases



Task 2

1. The occurrence of a disengagement in a mile can be approximated as a random variable with a Bernoulli distribution. (subject to the condition of maximum number of disengagement in a mile)

$$P(DPM|Cloudy) = \frac{P(Cloudy|DPM)P(DPM)}{P(Cloudy)} = 0.059 \quad \& \quad P(DPM|\text{Clear}) = \frac{P(Cloudy|DPM)P(DPM)}{P(Clear)} = 0.00051$$

   - Similarly, the probability of automatic disengagement per mile on a **cloudy** day is *0.0028* and on a **clear** day is *0.00026*.
   - Checked how likely it is to have over *150* disengagements in *10000* miles and which comes out to be nearly impossible.

2. Tested the hypothesis if the number of disengagements on a cloudy day is more than the number of disengagements on a clear day at the 0.05 significance level

$$H_o: \text{ Number of disengagement in cloudy} \leq \text{ Number of disengagement in clear}$$
$$H_a: \text{Number of disengagement in cloudy} > \text{Number of disengagement in clear}$$

   - Null hypothesis is rejected and concluded that the AV has more number of disengagements on a cloudy day.

Task 2

3. P(Reaction Time > 0.5s | Cloudy): 0.5390
   P(Reaction Time > 0.7s | Cloudy): 0.3854

4. Based on the assumption given, calculated the probability of an accident per mile involving an automatic disengagement which came out to be 0.00049.

$$P(acc/mile) = P(RT > 0.7s|Clear, DPM)P(DPM|Clear)P(Clear) + P(RT > 0.5s|Cloudy, DPM)P(DPM|Cloudy)P(Cloudy))$$

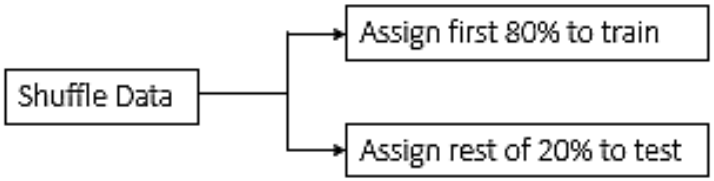
5.
```
P(Car Accident | Human Driver): 2e-06
P(Car Accident | AV): 0.0004968044193821021
The probability of a human driver causing a car accident is smaller than AVs.
```

- The ratio of probability of accident caused by AVs to the probability of accident caused by a human driver comes out to be 248.40 which it indicates that AVs is 2.5 times more likely to cause an accident than a human driver.
- Conclusion: AVs are currently not safer than human drivers.

Task 3

1. Grouping into 3 classes
   - Created new class column, and replaced each cause with the corresponding class label

2. Train Test Split



3. Training

Separate Data by Class C_k → $p(C_k) = \frac{\#\ of\ diseng\ with\ C_k}{Total\ \#\ of\ diseng}$ → $p(x_i|C_k) = \frac{\#\ of\ x_i\ in\ C_k}{Total\ \#\ of\ diseng\ with\ C_k}$

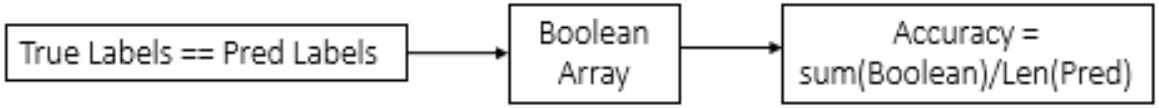
	Computer System	Controller	Perception System
Location	{'urban-street': 0.9426, 'highway': 0.0574}	{'urban-street': 1.0, 'highway': 0.0}	{'urban-street': 1.0, 'highway': 0.0}
TypeOfTrigger	{'automatic': 0.4959, 'manual': 0.5041}	{'automatic': 0.1514, 'manual': 0.8486}	{'automatic': 0.8351, 'manual': 0.1649}
Weather	{'cloudy': 0.3689, 'clear': 0.6311}	{'cloudy': 0.9965, 'clear': 0.0035}	{'cloudy': 1.0, 'clear': 0.0}

4. Testing/Prediction
   - Assumption: class conditional independence $p(x_i|x_{i+1}, x_{i+2}, .., x_n, C_k) = p(x_i|C_k)$
   - Compute probability using trained model
   - Predict label with maximum probability

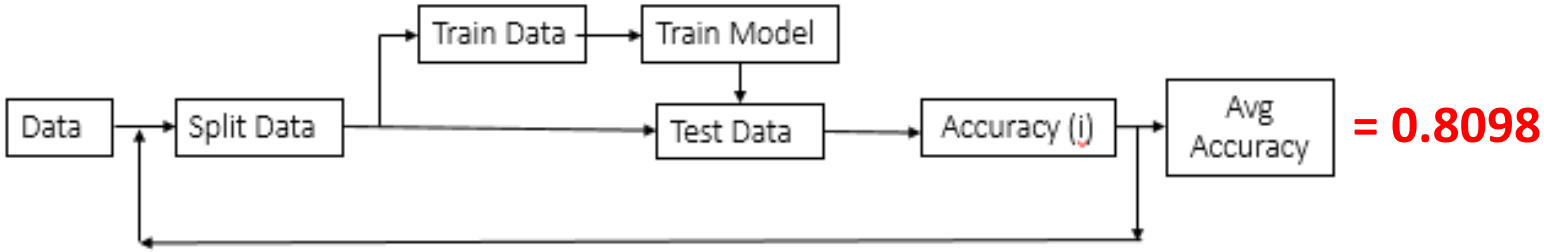
$$C^* = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Task 3

- Accuracy



5. Cross-Validation



6. Comparison with Chance
   - The chance level performance gives an accuracy of 0.335
   - Naive Bayes performs better than chance.
7. Assumptions in NB
   - Assume features are class conditional independent
   - The assumption is not valid for one of the cases and cannot be tested for some cases.
8. Improvements
   - The Naive Bayes model cannot be improved further except increasing training data size..
   - Class conditional independence may not hold for some cases and hence, some other model could be chosen which doesn't assume the same, for example, logistic regression..