Mini-Project 1

ECE/CS 498DS

Spring 2020

Abhinavi Madireddy (am49)

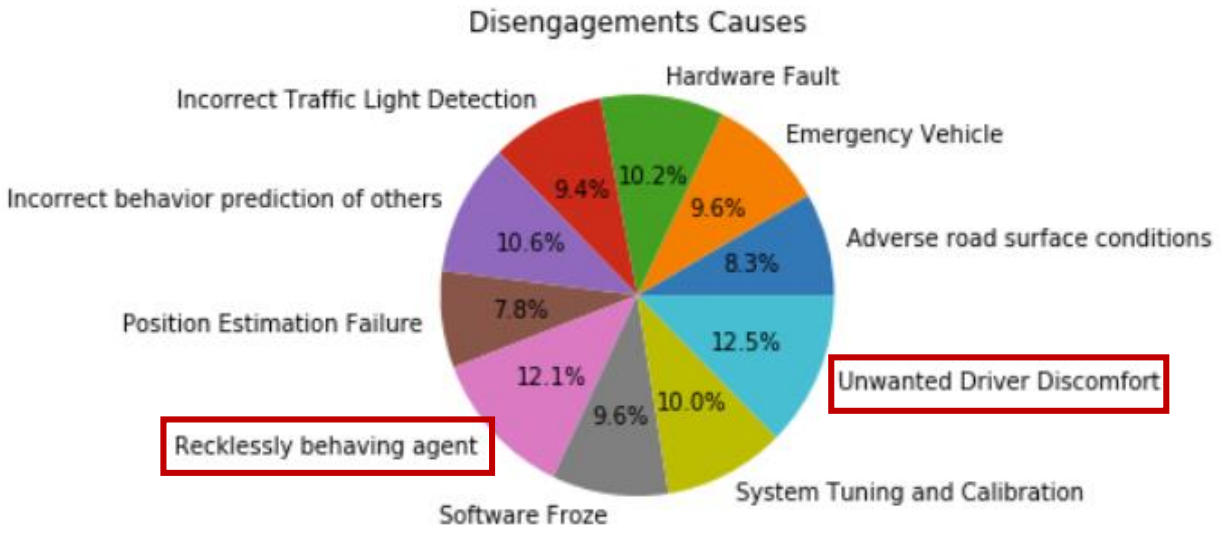
Yasaswi Boyapati (yasaswi2)

Rishabh Gupta (rishabh7)

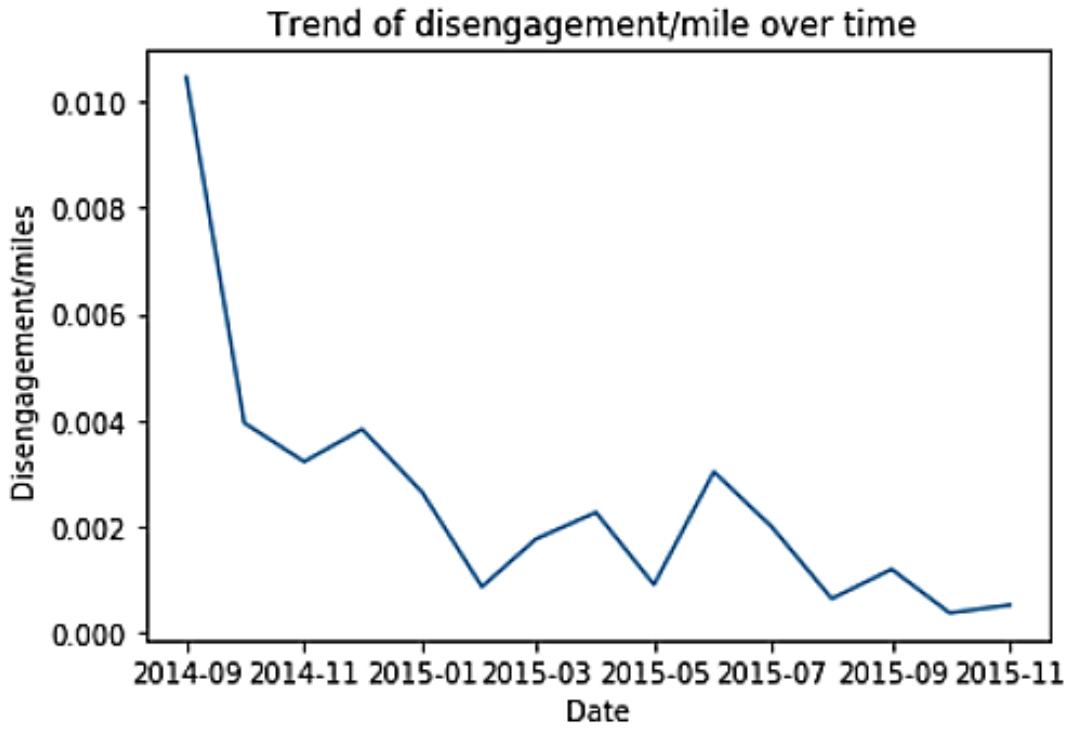
Task 0

Summarize	Answer
2(a) – Total Disengagements	1024
2(b) - # Unique Months	15
2(c) – Unique Locations	Urban-street, highway
2(d) - # Unique Causes	10
2(e) - # Rows with missing values	Reaction Time – 532 (RT for Manual)

Qn 3 - Causes of AV Disengagement



Qn 4 - Trend of disengagement/mile
Plot:



Task 1

Qn 1 – Interpreting Distributions

(a) Gaussian – Mean = Median = Mode

- Most of the data lies nearby the mean value.
- The pdf looks like bell shaped curve
- Defines Central Limit Theorem – states that the mean of the samples, collected from any data following any distribution, must follow normal distribution.

(b) Exponential – Skewed continuous distribution,

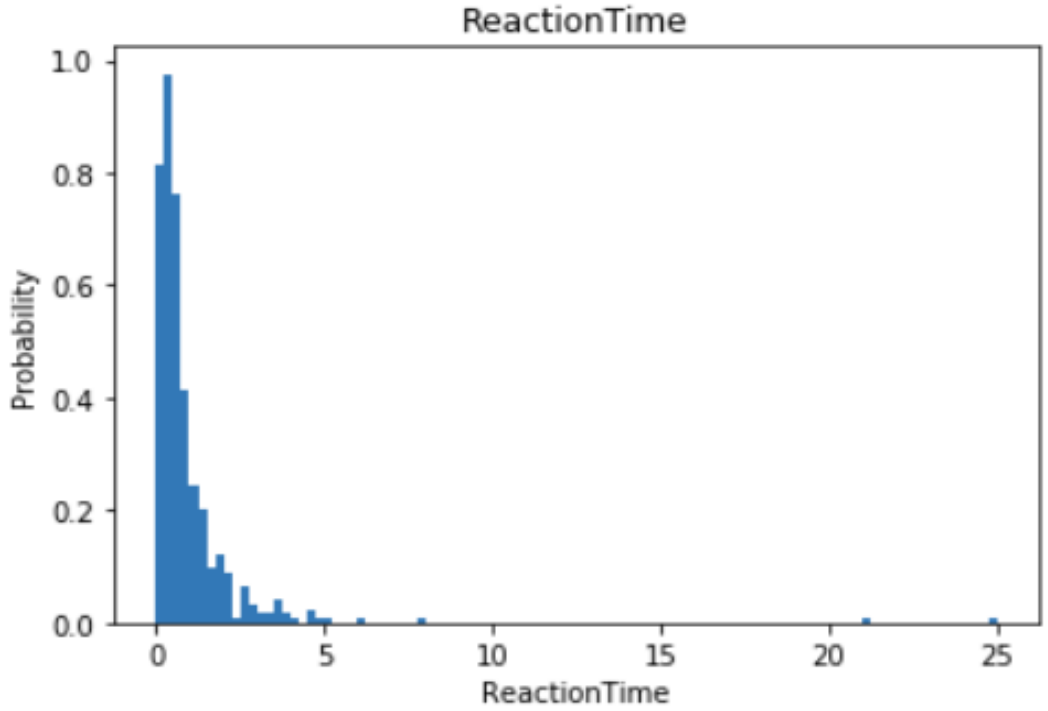
- Used for measuring the expected time of occurrence of any event
- Gradually decays with time
- Memoryless Property - Probability that certain event keep going on independent of start time

(c) Weibull –

- Models the life distribution, reliability and analysis of time to failure;
- Examples- warranty/guarantee analysis, failure of a machine/component.

Qn 2 – Probability distribution of reaction times

Plot:



What distribution does it fit and what is the significance?

- Looks like the distribution is Weibull distribution.
- Since here we are dealing with reaction time of disengagement and as Weibull distribution better describes the time to failure system, this statement supports that the data is closer to Weibull distribution.

Task 1

Qn 3 – Average Reaction times

(a) Over entire dataset: ~0.929s

(b) Over entire dataset per location:

- Highway : 1.48s
- Urban-Street : 0.92865s

Qn 4 – Hypothesis Testing

State the Null and Alternate Hypothesis

$H0 : \mu_{AV} = \mu_{non-AV} = 1.09$

$Ha : \mu_{AV} \neq \mu_{non-AV} \neq 1.09$

Statistic Value = Z-Test : -2.085

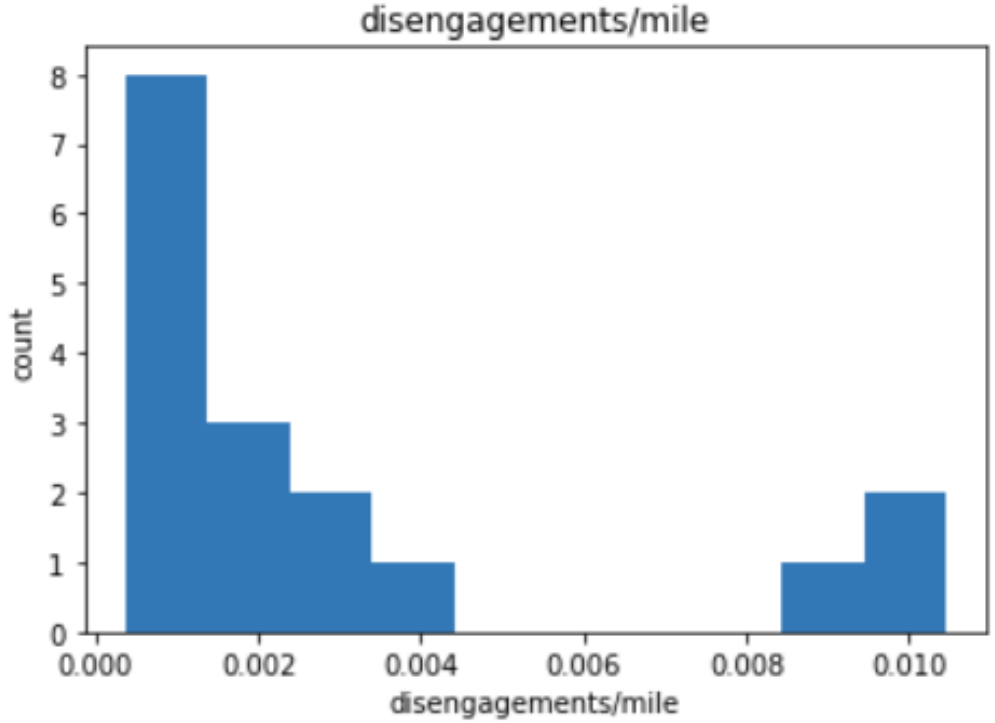
p-value : 0.037

Outcome of the hypothesis test:

- Rejected Null Hypothesis

Qn 5 – Probability distribution of disengagements/mile

Plot:



What distribution does it fit and what is the significance?

This histogram doesn't exactly fit any distribution but since the data used to build the histogram is quite restricted in terms of number of records, we believe that exponential distribution would be good fit to the data.

Task 2

Qn 1

(a) *Binomial* distribution

(b) Probability of disengagement/mile on a cloudy day:

~0.00590

(c) Probability of disengagement/mile on a clear day:

~0.00052

(d) Probability of automatic disengagement/mile

(i) on a cloudy day: ~0.00280

(ii) on a clear day: ~0.00026

(e) What approximation did you use? State the obtained probability in mathematical notation.

150 or more disengagement in 12000 miles is equivalent to 1 or more disengagements in 80 miles under cloudy conditions.

$$P_{12000}(k \geq 150) \approx P_{80}(k \geq 1) = 1 - P_{80}(k = 0)$$
$$= 1 - \binom{n}{0} p^0 (1-p)^{(80-0)} = 1 - (1-p)^{80}$$

= 0.377 (approx.)

Qn 2 – Hypothesis Testing Concepts

(a) All the real data ~ normal distribution and central limit theorem also favors the normal distribution; the hypothesis testing is performed using standard normal distribution.

- In hypothesis testing, the standard normal distribution tell us about the p-value based on Z statistics. Using these p-value of certain data point, we can estimate whether the data point is inside confidence interval or not.

(b) If the data does not support the alternative hypothesis, this does not mean that the null hypothesis is true. All it means is that the null hypothesis has not been disproven—hence the term "failure to reject." A "failure to reject" a hypothesis should not be confused with acceptance.

Task 2

Qn 3 – Z-test
State the Null and Alternate Hypothesis
Null: P_disengaged_mile_clear_day - P_disengaged_mile_cloudy_day >= 0
Alternate: P_disengaged_mile_clear_day - P_disengaged_mile_cloudy_day < 0
Statistic Value = -1.48
P-value=0.1388
Outcome of the hypothesis test: Reject Null Hypothesis
Conclusion: P_disengaged_mile_cloudy_day > P_disengaged_mile_clear_day

Qn 5 – Conditional Probability and Total Probability
(Write both the probability expression and the computed probability value)

P(Accident|Disengaged) = P(Accident|Disengaged,cloudy) * P(cloudy) + P(Accident|Disengaged, clear) * P(clear)

= 0.335 (approx.)

Qn 4 – Conditional Probability (Write both the probability expression and the computed probability value)
(a) P(RT>0.6|cloudy day, Auto Disengagement) = Auto Disengagement on cloudy RT>0.6/ Auto Disengagement on cloudy
= 0.473 (approx.)
(b) P(RT>0.9|clear day, Auto Disengagement) = Auto Disengagement on clear RT>0.9/ Auto Disengagement on clear
= 0.281 (approx.)

Task 2

Qn 6 – Comparing AVs to human drivers.
The probability of AVs causing accident is $\mathbf{4.25 \times 10^{-4}}$ which is around 100 times the probability of human driver ($\mathbf{2 \times 10^{-6}}$) causing a car accident. This means, AVs are more susceptible than human driver to cause a car accident.

Qn 7 – KS Test
State the Null and Alternate Hypothesis
Null: Both the distributions are same
Alternate: Both the distributions are different

Statistic Value = 0.056
P value = 0.95

Outcome of the hypothesis test: Fail to reject null hypothesis.
Conclusion: Both the samples follows same distribution.

Even though, the both cloudy and clear reaction times are from same distribution, we cannot say anything about the impact of weather conditions on disengagement reaction time. However, if we can calculate the Z statistics that whether the mean of cloudy reaction time is equal to the mean of clear reaction time or not. This means the null and alternate hypothesis will be:

Null Hypothesis : $\mu_{cloudy\_rt} = \mu_{clear\_rt}$

Alterante Hypothesis : $\mu_{cloudy\_rt} \neq \mu_{clear\_rt}$

Task 3

Qn 3 – Conditional Probability Tables for NB classifier

		Location		Weather		TypeOfTrigger	
	Class	highway	urban-street	clear	cloudy	automatic	manual
Class							
Computer System	0.300000	0.069106	0.930894	0.617886	0.382114	0.483740	0.516260
Controller	0.365854	0.000000	1.000000	0.003333	0.996667	0.113333	0.886667
Perception System	0.334146	0.000000	1.000000	0.000000	1.000000	0.832117	0.167883

Qn 4 – NB Classifier Accuracy: 75.9%

Qn 5 – NB Cross Validation Accuracy: 79.6%

Task 3

Qn 6 – Is your NB model doing better than chance? Explain.
Yes, NB model is doing great. The average accuracy is almost **80%**, which is far better than chance. Since earlier before creating Naive Bayes model, P(Class = Computer System) ~ 0.287, P(Class = Controller) ~ 0.3609, P(Class = Perception System) ~ 0.3512. Based on these probability, the probability of controller class was higher. This mean the accuracy of model was **36.09%**. However, now with NB model, the accuracy improved to 80% (approx.).

Qn 7 – Assumptions in NB in this context of AV safety analysis.
All the features i.e., location, weather and type of trigger are conditionally independent.
This assumption may not be realistic since there might be some relationship between type of trigger vs weather condition and type of trigger vs location; since, there are high chances of disengagements per mile in urban-street than highway and during cloudy than clear weather condition.

Qn 8 – Possible improvements to increase classification accuracy.
Since, we assumed about independency of the features, which might not be completely right. We might be able to improve the accuracy of the model by taking the dependency among the features into account using Bayesian Network.

Insights on AV safety

- **List some insights on AV safety that you have gained by performing data analysis on the CA DMV dataset**
  - Probability of disengagement causes are very close to each other.
  - The trend of disengagement/mile decreased over time
  - Average reaction time was more in highway than urban-street.
  - The probability of automatic disengagement per mile on a cloudy day is 10 times higher than on a clear day
  - AVs are more susceptible than human driver to cause a car accident.
- **Would you ride in an AV based on the data you have analyzed**?
  - No, since the chance of accident in AVs vs human driver is 100 times higher.
- **What do you think about the future of AVs and how soon they will be deployed?**
  - We think with few improvements, AVs should be deployed in the next 5-10 years. Still, a lot of work to improve the algorithm in making real time decisions is needed.
- **What would you change about the MP? What other analysis would you have done?**
  - The problem in the MP were quite interesting and insightful.
  - We could have developed the model to predict the actual causes instead of class of causes. For that we need more data.
  - We could have incorporated/checked the dependency among the features- location, weather and type of trigger to develop even better Naïve Bayes Model.

Individual Contributions

Abhinavi Madireddy – Equal Contribution

Yasaswi Boyapati – Equal Contribution

Rishabh Gupta – Equal Contribution

We individually coded for all the questions and then discussed for each question about the correct way to analyse the problems.

Thank You!