Mini-Project 1

ECE/CS 498DS

Spring 2020

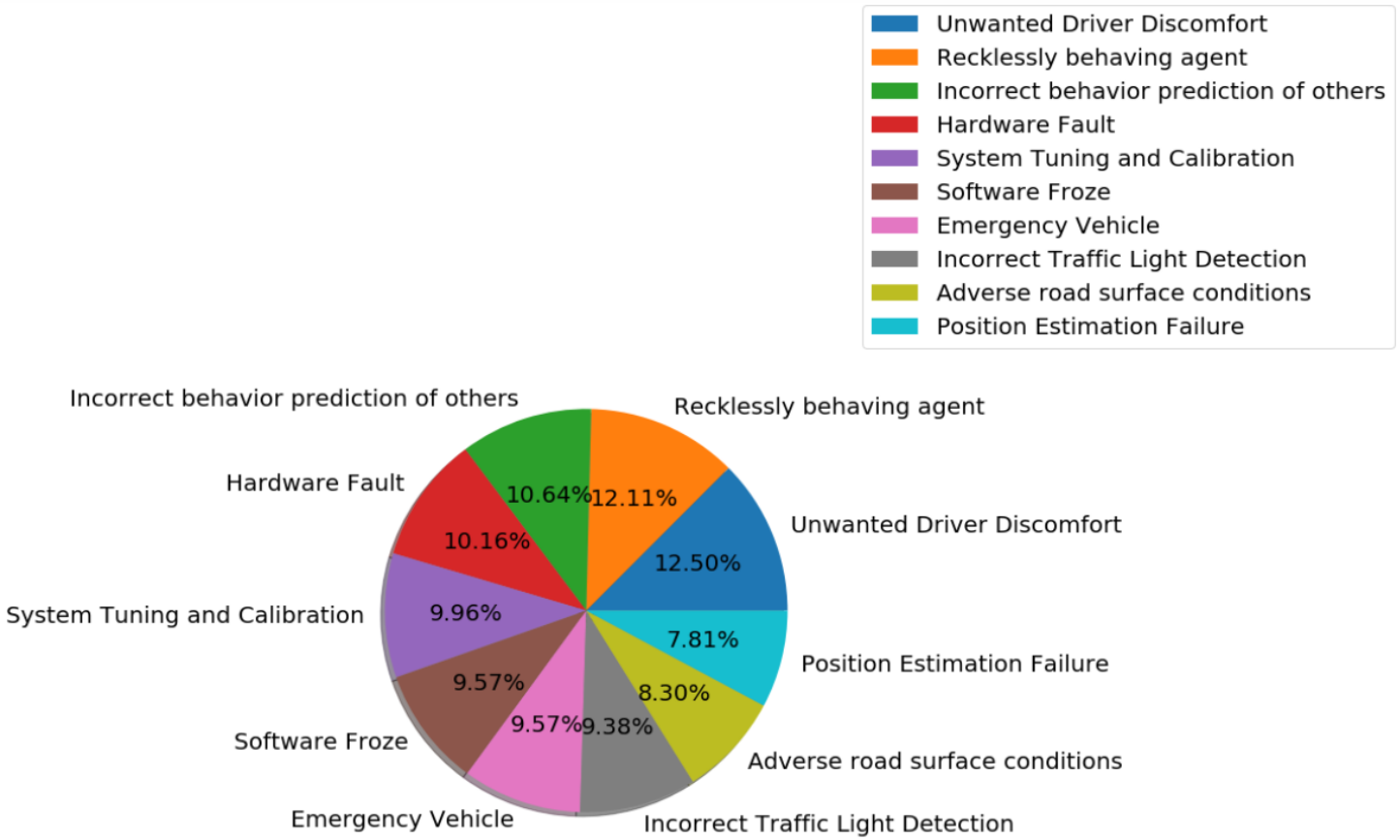
Meghna Shrivastava (meghna3), Mohit Jain (mohitj2), Tafseer Khan (tafseer2)

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	ReactionTime-532

Task 0

Question 3:

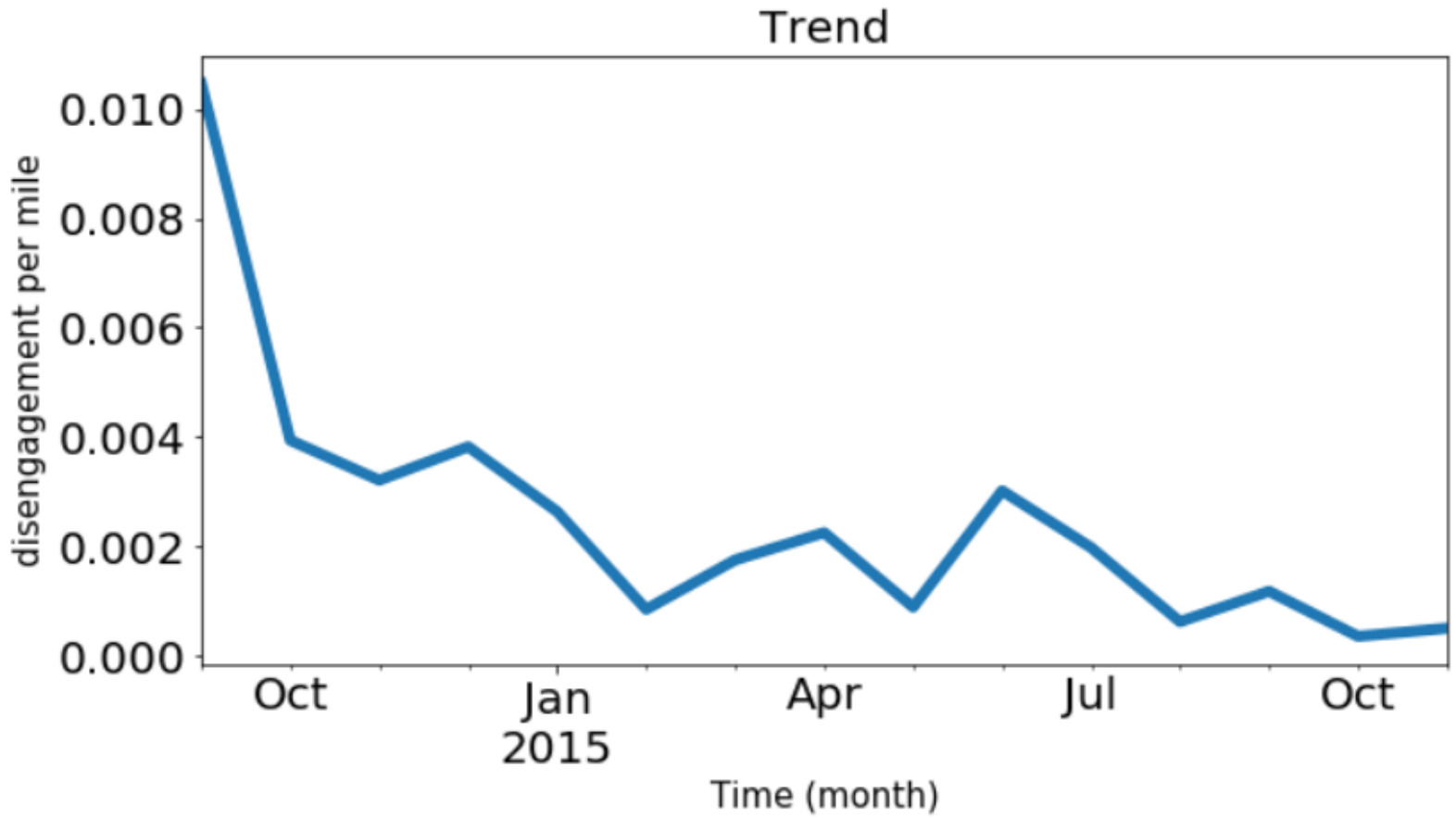


According to the pie chart the top 2 leading causes of disengagement are Unwanted Driver Discomfort & Recklessly behaving agent

Task 0

Question 4 - Trend of disengagement/mile

Plot:



Task 1

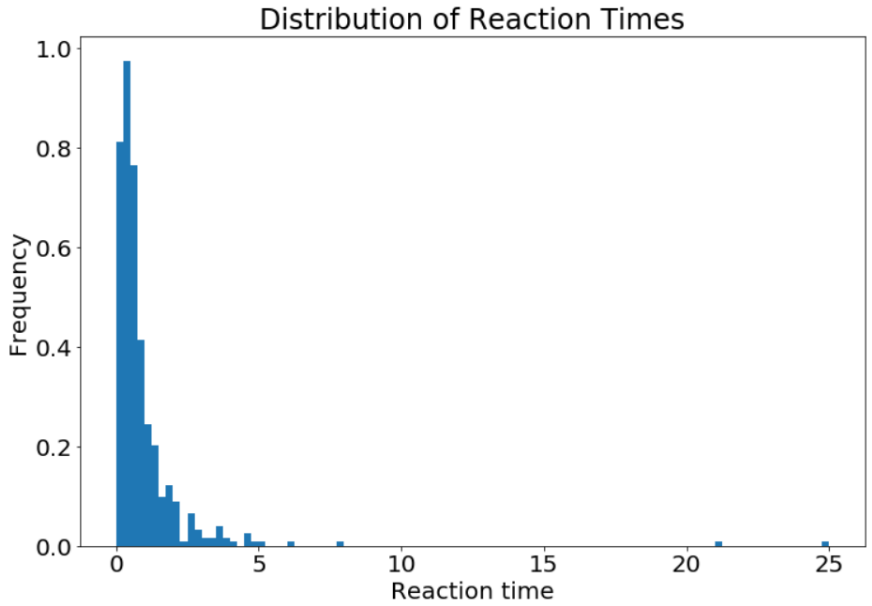
Qn 1 – Interpreting Distributions

(a) Gaussian –
This distribution also known as the “Bell Curve”. And because of the following features it is highly appreciated by the data scientists.
The mean, median and mode of normal distribution are equal plus it is symmetric around the mean.
The curve is dependent on the mean and standard distribution of their data and it is very dense at the center and less dense at the tails.
Appoximately 95% of the area of the curve is within 2 standard deviations of the mean
(b) Exponential –
The exponential distribution describes the amount of time between occurrences.
E[X] is given by 1/ λ where λ Exponential Distribution .
Which means that as λ gets larger the less is the time between occurrences.
For Poisson equation, Exponential Distribution is useful to model the random arrival pattern
(c) Weibull –
It is widely used in life like systems for the data analysis.
It has 2 parameters, 1- Beta as shape parameter and second is N a scale parameter.
If beta is less than 1, the probailty density tends to infinity at time ~zero
If beta is equal to 1 the graph the failure rate is fairly constant.
If beta is greater than 1 the failures rate increases as time increases.

Task 1

Qn 2 – Probability distribution of reaction times

Plot:



What distribution does it fit and what is the significance?

According to the plot it can be inferred that the distribution follows weibul distribution which shows that the probability of the reaction time being high for human beings is very low.

Qn 3 – Average Reaction times

(a) Over entire dataset: 0.9297703252032521

(b) Over entire dataset per location:

Location	ReactionTime
highway	1.48000
urban-street	0.92865

Task 1

Qn 4 – Hypothesis Testing

State the Null and Alternate Hypothesis
Null Hypothesis: “Reaction time for humans in AV cars is not different from non-AV cars.”
Alternate Hypothesis: “Reaction time for humans in AV cars is different from non-AV cars.”

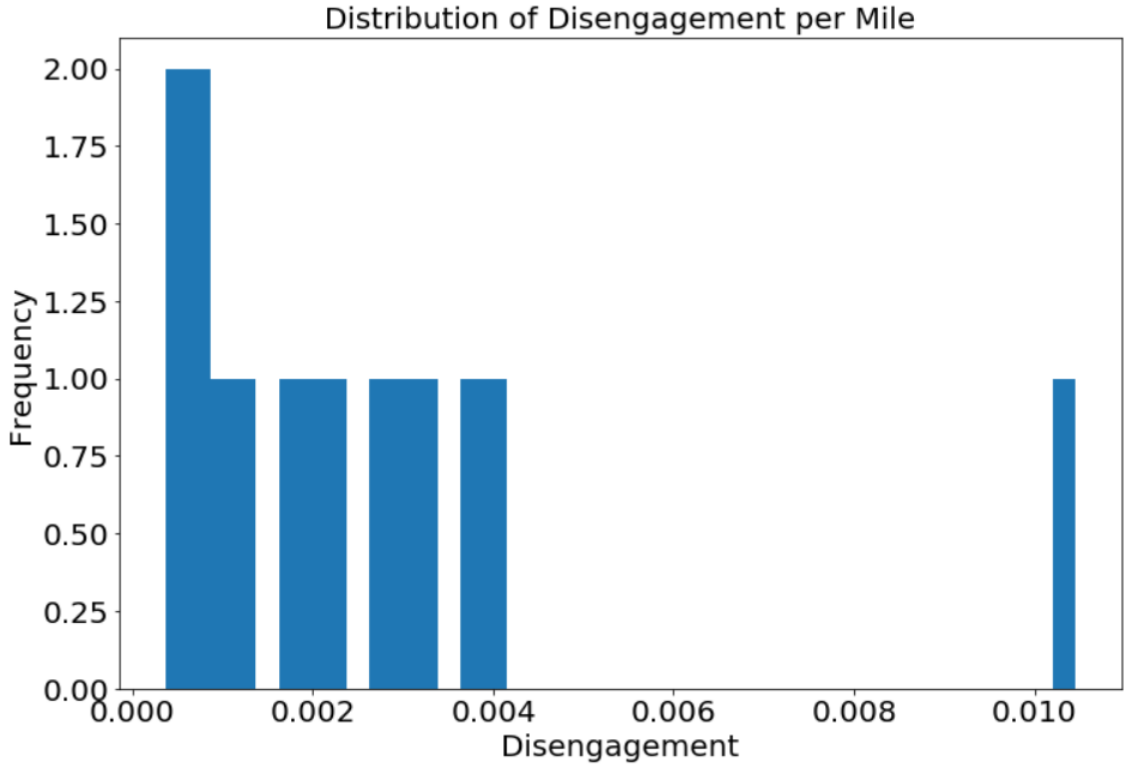
Significance level = 0.05
P-value= 0.036

Outcome of the hypothesis test: Null hypothesis rejected.
Reaction time for humans in AV cars is different from non-AV cars

Task 1

Qn 5 – Probability distribution of disengagements/mile

Plot:



The distribution mostly fits an exponential distribution. The disengagement per mile is decreasing continuously expect for a few values which are rising. These values can be considered as outliers. We will be able to have a much better idea of the fit with a larger dataset.

Task 2

Qn 1

(a) Bernoulli distribution

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

```
P(dpm|cloudy): 0.00590255677552725
```

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

```
P(dpm|clear):  0.0005195663748517999
```

(d) Probability of automatic disengagement/mile

(i) on a cloudy day: `P(auto,dpm|cloudy): 0.0028063653172267287`

(ii) on a clear day: `P(auto,dpm|clear): 0.00026390673008345393`

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

We can assume that the distribution is normal.

$$\mu = np = 12000 \times P(dpm|cloudy)$$

$$\sigma^2 = np(1-p) = 12000 \times P(dpm|cloudy)(1 - P(dpm|cloudy))$$

$$Z_n = (X_n - \mu)/\sigma$$

```
Mean= 70.830681306327
Std= 8.391221555200836
z-score= 9.43477873547556
p-value= 1.9591016830538447e-21
```

Task 2

Qn 2 – Hypothesis Testing Concepts

(a) In Hypothesis testing, the hypothesis tests of a population mean is performed using the normal distribution. It is necessary to generalize the hypothesis test results to a population. Also, the normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large. Normal Distribution in hypothesis testing basically helps in determining if the sample that has been tested falls in the critical areas. If that's the case, then according to the concept of Hypothesis testing, null hypothesis gets rejected and alternative testing gets considered. The 'two-tailed' test is derived from testing the area under both tails of a normal distribution too. It helps in giving an estimate of what is possible. Assuming a normal distribution also lets us determine how meaningful the result we observe in a study is. For eg: The higher or lower the z-score in Hypothesis testing, the more unlikely the result is to happen by chance and the more likely the result is meaningful.

(b) In the hypothesis testing, both the Ho and Ha are assumed to be two sides of an extreme i.e either having Ho or the Ha probability. If null hypothesis means that there is no variation perhaps the statistical significance in the set of observations considered then rejecting this hypothesis eventually signifies the only other possibility left i.e Ha.

Task 2

Qn 3 – Z-test

State the Null and Alternate Hypothesis

H_o : Number of disengagement in cloudy $\leq$ Number of disengagement in clear

H_a : Number of disengagement in cloudy $>$ Number of disengagement in clear

Statistic Value = `38.1986243877602`

P-value= 0

Outcome of the hypothesis test: Reject null hypothesis

Conclusion:

Since p-value is low, we can reject the null hypothesis and conclude that the number of disengagements on a cloudy are more.

Qn 4 – Conditional Probability

(Write both the probability expression and the computed probability value)

(a) `P(Reaction Time > 0.6s | Cloudy): 0.473551637279597`

(b) `P(Reaction Time > 0.9s | Clear): 0.28125`

Qn 5 – Conditional Probability and Total Probability

(Write both the probability expression and the computed probability value)

$$P(acc/mile) = P(rt > 0.9s|clear, dpm)P(dpm|clear)P(clear) + P(rt > 0.6s|cloudy, dpm)P(dpm|cloudy)P(cloudy))$$

`P(acc/mile): 0.0005621213350777568`

Task 2

Qn 6 – Comparing AVs to human drivers

```
P(Accident|Human): 2e-06
P(Accident|AV): 0.00056212133507775 68
The probability of a human driver causing car accident is lesser than AVs.
```

Qn 7 – KS Test

State the Null and Alternate Hypothesis

Null_H: Both distributions are of same type
Alternate_H: The distributions are not same

Statistic Value = 0.05622900923593619

P value = 0.9534988141679469

Outcome of the hypothesis test & Conclusion:

Since the p-value is so high we accept the null hypothesis and conclude that the distributions are similar which signifies that the weather being cloudy or clear has no effect on the reaction time of a person. Also, from the test results obtained, the disengagement reaction time when the weather is clear is more as compared to when the weather is cloudy.

Task 3

Qn 3 – Conditional Probability Tables for NB classifier

Priors:

Class	
Computer System	0.296875
Controller	0.352539
Perception System	0.350586

Class	TypeOfTrigger	
Computer System	automatic	0.483553
	manual	0.516447
Controller	automatic	0.132964
	manual	0.867036
Perception System	automatic	0.830084
	manual	0.169916

Class	Weather	
Computer System	clear	0.618421
	cloudy	0.381579
Controller	clear	0.002770
	cloudy	0.997230
Perception System	cloudy	1.000000

Class	Location	
Computer System	highway	0.065789
	urban-street	0.934211
Controller	urban-street	1.000000
Perception System	urban-street	1.000000

Qn 4 – NB Classifier Accuracy: 54.146341463414636

Qn 5 – NB Cross Validation Accuracy: CV-score: 0.5942435338266052

```
Accuracy of the model is:  75.1219512195122
Accuracy of the model is:  52.6829268292683
Accuracy of the model is:  53.170731707317074
Accuracy of the model is:  72.6829268292683
Accuracy of the model is:  58.536585365853654
Average Accuracy of 5 cases is  62.4390243902439
```

Task 3

Qn 6 – Is your NB model doing better than chance? Explain.

Since the model accuracy is above 50% on average, we can say that our model is doing better than chance. However there are instances where the accuracy drops very close to 50% due to which we cannot rely on our model completely.

Qn 7 – Assumptions in NB in this context of AV safety analysis

NB makes naive assumption that the attributes of the dataset are conditionally independent of eachother. The attributes in this case are the location, weather and type of trigger.The class labels are controller, perception system and computer system. These assumptions are not completely realistic in real world situations. No assumption can be completely realistic in real world scenarious but NB still provides a decent classification even with the assumptions.

Qn 8 – Possible improvements to increase classification accuracy

Keeping NB as the base model we can look at more sophisticated techniques like Bayesian Networks, HMM, clustering, etc which do not make assumptions of conditional independence and compare accuracy of the predictions with NB model.

Insights on AV safety

- List some insights on AV safety that you have gained by performing data analysis on the CA DMV dataset

```
P(Accident|Human): 2e-06
P(Accident|AV): 0.0005621213350777568
The probability of a human driver causing car accident is lesser than AVs.
Ks_2sampResult(statistic=0.05622900923593619, pvalue=0.9534988141679469)
```

Since the p-value is so high we accept the null hypothesis and conclude that the distributions are similar which signifies that the weather being cloudy or clear has no effect on the reaction time of a person. Also, from the test results obtained, the disengagement reaction time when the weather is clear is more as compared to when the weather is cloudy.

- Would you ride in an AV based on the data you have analyzed?

Based on the analysis we can conclude that AVs are not safe. We would not ride in an AV as of yet.

- What do you think about the future of AVs and how soon they will be deployed?

AVs are improving continuously as more and more data is being collected. Companies like Tesla and Waymo have already developed AV systems that are almost perfect. With on-going cutting-edge research in the fields of machine learning, deep learning & computer vision, the future of AVs is bright. Tesla and Waymo have already deployed AVs and in a few more years AVs can be seen all over the world.

- What would you change about the MP? What other analysis would you have done?

The only thing I would like to change about is the dataset. I feel it is too small to come up with actionable insights. We are interested in applying other classification techniques like Bayesian networks, logistic regression, SVM, decision trees and compare the accuracy of the models.

Individual Contributions

Meghna – Initial data analysis, NB model CV test (Task 0, Task 2, Task 3)

Mohit – Initial data analysis, plotting of graphs, insights, NB model implementation (Task 0, Task 1, Task 3)

Tafseer – Initial data analysis, hypothesis testing, probability calculations, NB model insights (Task 0, Task 2, Task 3)