Mini-Project 1

ECE/CS 498DS
Spring 2020

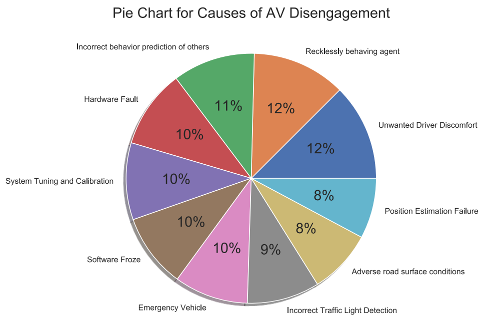
Chuhao Feng (chuhaof2), Boyang Zhou (boyangz3), Mengxuan Yu (my13)

All registered

Task 0

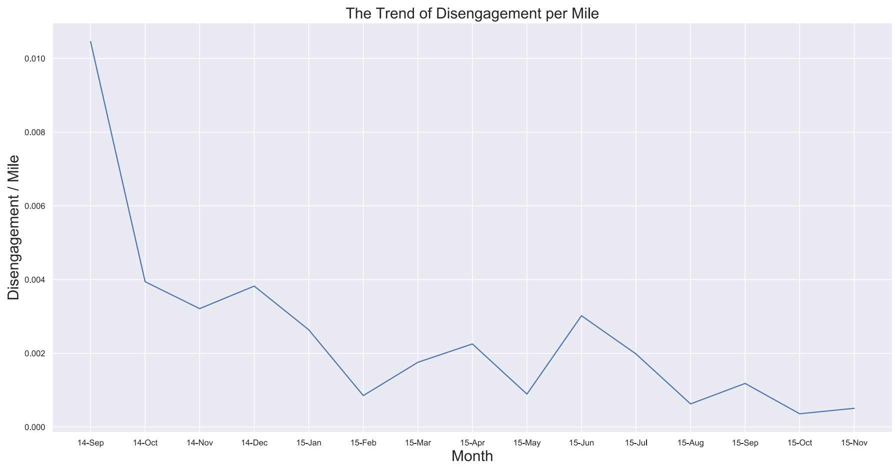
Summarize	Answer
2(a) – Total Disengagements	1024
2(b) - # Unique Months	15
2(c) – Unique Locations	['urban-street' 'highway']
2(d) - # Unique Causes	10
2(e) - # Rows with missing values	532

Qn 3 - Causes of AV Disengagement
Pie chart:



Top 2 causes of disengagement:
Recklessly behaving agent
Unwanted Driver Discomfort

Qn 4 - Trend of disengagement/mile
Plot:



Task 1

Qn 1 – Interpreting Distributions

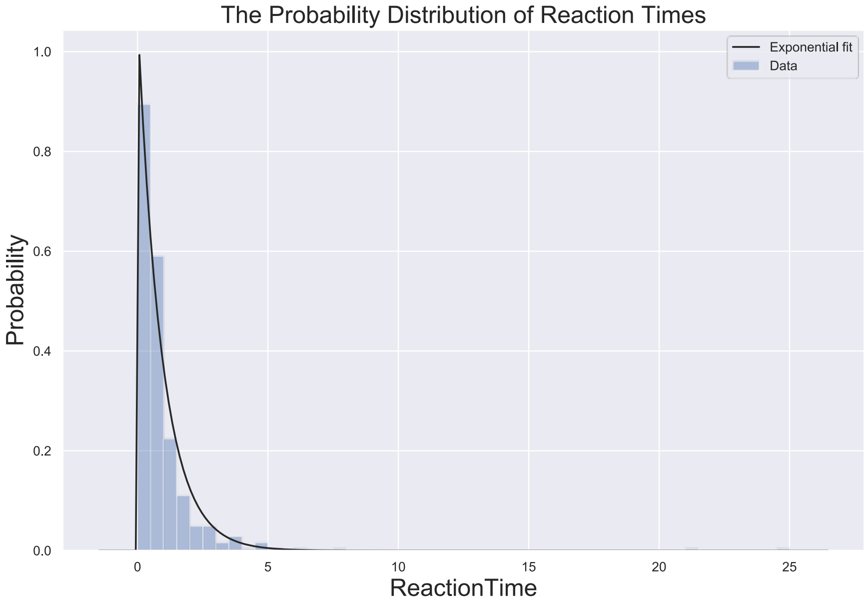
(a) Gaussian – Gaussian distribution tells us that data near the mean occurs more frequently than data far from the mean.
(b) Exponential – Exponential distribution tells us that the probability for sample to occur decreases exponentially as time goes by, often used to model the time elapsed between events.
(c) Weibull – Weibull distribution is often used to model lifetime of components or time to failure, telling us how the probability for failure behaves as time goes by.

Qn 3 – Average Reaction times

(a) Over entire dataset: 0.929770
(b) Over entire dataset per location:

highway	1.48000
urban-street	0.92865

Qn 2 – Probability distribution of reaction times
Plot:



What distribution does it fit and what is the significance?

This distribution fits exponential distribution. The fit distribution signifies that the probability of reaction time decreases exponentially as reaction time goes up. In other words, human driver tends to react immediately.

Task 1

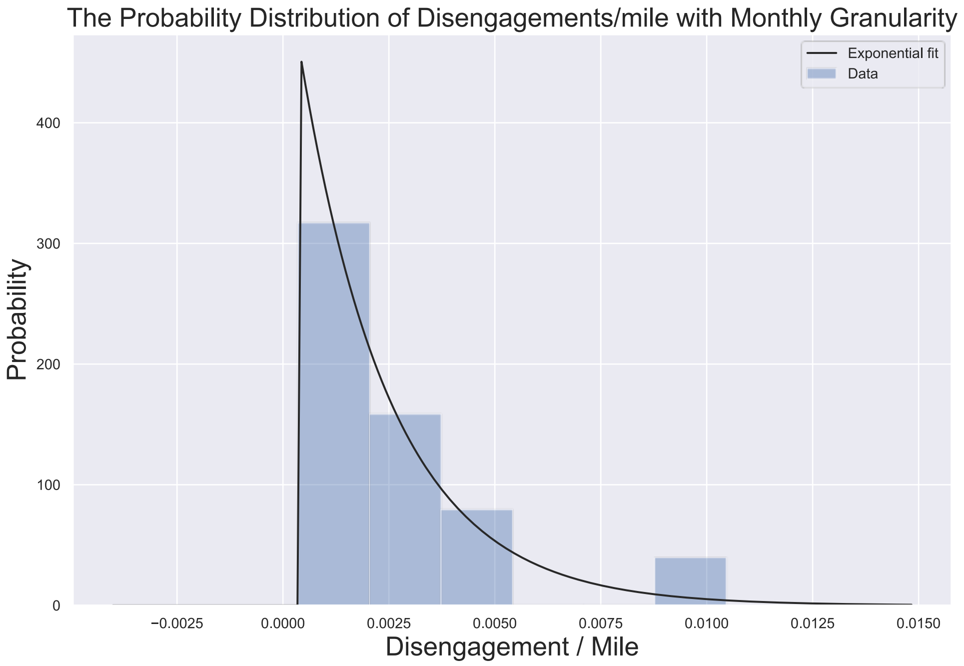
Qn 4 – Hypothesis Testing

State the Null and Alternate Hypothesis
Null hypothesis: The mean reaction time for humans in AV cars is the same as that in non-AV cars, i.e. $\mu = 1.09$.
Alternate hypothesis: The mean reaction time for humans in AV cars is different from that in non-AV cars, i.e. $\mu \neq 1.09$.

Statistic Value = -2.098607967964133 < -1.645

Outcome of the hypothesis test
Reject null hypothesis.
The mean reaction time for humans in AV cars is **different** from that in non-AV cars.

Qn 5 – Probability distribution of disengagements/mile
Plot:



What distribution does it fit and what is the significance?

This distribution fits exponential distribution. The distribution tells us that AVs perform relatively good, and disengagement/mile has higher probability to occur when it is smaller. Besides, high disengagement/mile seems impossible to happen, because probability for higher disengagement/mile decreases exponentially, which is good.

Task 2

Qn 1

(a) _Bernoulli_ distribution

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

p(D|C)

=# of disengagement in cloudy day/ (total miles*P(C))

=835/(505229*0.28)=0.005902556775527249

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

p(D|C')

=# of disengagement in clear day/ (total miles*P(C'))

=189/(505229*0.72)=0.0005195663748517998

(d) Probability of automatic disengagement/mile

(i) on a cloudy day: p(D|A,C)= 0.0028063653172267283

(ii) on a clear day: p(D|A,C')= 0.0002639067300834539

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

Using Poisson distribution:

P(X=150)=p(λ=0.0059*12000) (k>=150)=1.1852253231672e-12

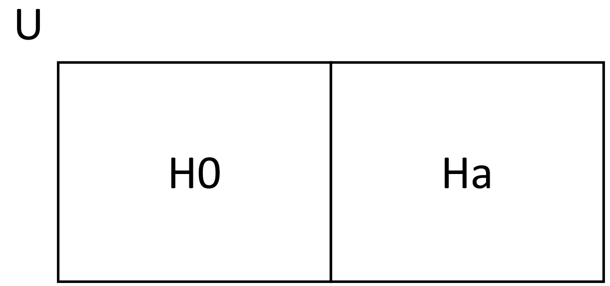
Qn 2 – Hypothesis Testing Concepts

(a)

The normal distribution represents the distribution of observed data If our Null Hypothesis is true.

(b)

Yes, according to the definition of Null and Alternative Hypothesis, Alternative Hypothesis is the opposite of the Null Hypothesis, which means they are mutually excusive and cover the universe set. Thus, rejecting H0 can mean accepting Ha. However, even if we can reject H0, it does not mean H0 is impossible to happen.



Task 2

Qn 3 – Z-test
State the Null and Alternate Hypothesis
H0: p(D|C)-p(D|C')<=0
Ha: p(D|C)-p(D|C')>0
Statistic Value = Z = -38.1986243877602

P-value = 0 (too small to calculate)

Outcome of the hypothesis test:
Z= -38.1986243877602 < -1.645
H0 Reject at α=0.05
Ha true
Conclusion: Cloudy days has larger possibility to have disengagement

Qn 4 – Conditional Probability
(Write both the probability expression and the computed probability value)

(a) P(reaction>0.6|cloudy)=(number of automatic disengagement with reaction time>0.6 on cloudy day)/(number of automatic disengagement on cloudy day)= 0.473551637279597
(b) P(reaction>0.9|clear)=(number of automatic disengagement with reaction time>0.9 on clear day)/(number of automatic disengagement on clear day)= 0.28125

Qn 5 – Conditional Probability and Total Probability
(Write both the probability expression and the computed probability value)
P(A)=P(Accident,Auto,Cloudy)+ P(Accident,Auto,Clear)
=(P(Accidnet|Auto,Cloudy)* P(auto,cloudy)+P(Accident|Auto,Clear)* P(auto,clear)
=(P(Accidnet|Auto,Cloudy)* (number(auto disengagement,cloudy)/total cloudy miles) +P(Accident|Auto,Clear)* (number(auto disengagement,clear)/total clear miles)
=0.0014031826586133641

Qn 6 – Comparing AVs to human drivers

Task 2

The probability of an accident happens of AV is 0.0014031826586133641
which is much greater than the human driver accident probability 2e-6.
This result shows that when AV disengagement happens automaticly it has a very high probability to result in an accident and the AV is not safe for the passengers. This reminds us that we should reduce the chance that an automatic disengagement happens in our development of AV system. Because once it happens, it has a great probability to cause an accident. And it also remind us to design a better system (such as an automatic disengagement prediction system) which can reduce the reaction time for the driver to take action to the automatic disengagement.

Qn 7 – KS Test
State the Null and Alternate Hypothesis
Null hypothesis: Samples of disengagement reaction time in cloudy days and samples of disengagement reaction time in clear days are drawn from the **same** continuous distribution, i.e. distribution of disengagement reaction time when the weather is cloudy **is the same as** distribution of disengagement reaction time when the weather is clear.
Alternative hypothesis: Samples of disengagement reaction time in cloudy days and samples of disengagement reaction time in clear days are drawn from **different** continuous distributions, i.e. distribution of disengagement reaction time when the weather is cloudy **is different than** distribution of disengagement reaction time when the weather is clear.

Statistic Value = 0.05171413502109701
P value = 0.9835003389445768 > α = 0.1

Outcome of the hypothesis test: Fail to reject null hypothesis
Conclusion: Distribution of disengagement reaction time when the weather is cloudy **is the same as** distribution of disengagement reaction time when the weather is clear, i.e. weather conditions have no impact on disengagement reaction time.

Task 3

Qn 3 – Conditional Probability Tables for NB classifier

Location	Class	P(Location|Class)
highway	Computer System	0.06867
highway	Controller	0.00000
highway	Perception System	0.00000
urban-street	Computer System	0.93133
urban-street	Controller	1.00000
urban-street	Perception System	1.00000

Weather	Class	P(Weather|Class)
Cloudy	Computer System	0.381974
Cloudy	Controller	0.996587
Cloudy	Perception System	1.000000
Clear	Computer System	0.618026
Clear	Controller	0.003413
Clear	Perception System	0.000000

TypeOfTrigger	Class	P(TypeOfTrigger|Class)
automatic	Computer System	0.463519
automatic	Controller	0.133106
automatic	Perception System	0.829352
manual	Computer System	0.536481
manual	Controller	0.866894
manual	Perception System	0.170648

Qn 4 – NB Classifier Accuracy: 0.7853658536585366

Qn 5 – NB Cross Validation Accuracy: 0.7873170731707317

Note that the values listed in this page vary every time the NB classifier is run.

Task 3

Qn 6 – Is your NB model doing better than chance? Explain.
The NB model is doing better than chance. If the model is doing by chance, the average accuracy should be 1/3 , because there are three types of class in total. Our NB model reaches an average accuracy that is larger than 0.6, showing that it is doing better than chance.

Qn 7 – Assumptions in NB in this context of AV safety analysis
Generally speaking, NB assumes that the features are class conditionally independent, i.e. independent given/conditioning on the class. In the context of this problem, features, Location, Weather, and TypeOfTrigger, are assumed to be class conditionally independent.

Based on the common sense, we do not think the assumptions are realistic. For example, the type of trigger may be affected by the type of location. We think that a manual trigger may have a higer probability to happen when the AV drives on urban streets than when the AV drives on a highway, because many people walk around the urban streets and the human driver may be more nervous about the behavior of the AV, which leads to a higher probability that the human driver takes control of the vehicle manually.

Qn 8 – Possible improvements to increase classification accuracy
We think that some improvements can be gained in classification accuracy.
For example, we can use local semantics to simplify the conditional probabilities after we apply the chain rule instead of simply assuming all the features are class conditionally independent. Making use of local semantics allows the model to take the relationship between features into account, which may, hopefully, improve the classification accuracy.

Insights on AV safety

- List some insights on AV safety that you have gained by performing data analysis on the CA DMV dataset
- The Naïve Bayes model tell us that the computer system is the most common cause of the disengagement. (6 in 8 cases) This means we should develop a more reliable computer system or a more advanced AI to increase the AV safety.
- The high accident probability when auto disengagement happens reminds us that we should reduce the chance that an automatic disengagement happens in our development of AV system. Because once it happens, it has a great probability to cause an accident. And it also remind us to design a better system (such as an automatic disengagement prediction system) which can reduce the reaction time for the drive to take action to the automatic disengagement.

- Would you ride in an AV based on the data you have analyzed?
- No, I won’t ride the AV. This result shows that when AV disengagement happens automatically it has a very high probability to result in an accident and the AV is not safe for the passengers.

- What do you think about the future of AVs and how soon they will be deployed?
- I think AV will be soon popularized since the AI used for AV is not very complex. Thus, it is easy to improve the AV to be safer than human driver. Though some factor may effect the disengagement rate. We can build some complementary device for in the street (like the traffic light or road line can send the message directly into AV). It won’t cost too much, but will bring great market for AV industrial.

- What would you change about the MP? What other analysis would you have done?
- In the MP, we use the Naïve Bayes Model to make prediction of the cause of fault according to the features given. But we can not be sure about the independence of the given features. So we want to change the prediction method to the Bayes Network. This method will include the dependence to our prediction system. Then we can compare the result of the two method. This will give us some insight about the relation between the features. To get a quantity description of the dependence of the features (how much a feature affects one another), we can use the linear regression to explore the relations between the features. Finally we will have an idea on how the features work together to affect the cause of fault.

Individual Contributions

Mengxuan Yu
- Task 1
- Task 2 Q1-Q3
- Insights Q3

Boyang Zhou
- Task 0
- Task 2 Q4-Q6
- Task 3 Q1-Q5
- Insights Q1 Q2 Q4

Chuhao Feng
- Task 2 Q7
- Task 3 Q6-Q8
- Insight Q4
- Polish Task 0
- Polish Task 1
- Polish Task 3