Pawin Piemthai - 5931037621

T1

$$P(y_2, y_1, y_0 \mid \alpha) = P(y_2 \mid y_1, y_0, \alpha) \cdot P(y_1, y_0 \mid \alpha)$$

$$P(y_1, y_0 \mid \alpha) = P(y_1 \mid y_0, \alpha) \cdot P(y_0 \mid \alpha)$$

Hence, $P(y_2, y_1, y_0 \mid \alpha) = P(y_2 \mid y_1, y_0, \alpha) \cdot P(y_1 \mid y_0, \alpha) \cdot P(y_0 \mid \alpha)$

Since, y_2 and y_0 are independent.

So, $P(y_2 \mid y_1, y_0, \alpha) = P(y_2 \mid y_1, \alpha)$

$$P(y_2, y_1, y_0 \mid \alpha) = P(y_2 \mid y_1, \alpha) \cdot P(y_1 \mid y_0, \alpha) \cdot P(y_0 \mid \alpha)$$

$$P(y_2 \mid y_1, \alpha) = P(w_1 = y_2 - \alpha y_1)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_2 - \alpha y_1 - 0)^2}{2\sigma^2}}$$

$$P(y_1 \mid y_0, \alpha) = P(w_0 = y_1 - \alpha y_0)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_1 - \alpha y_0 - 0)^2}{2\sigma^2}}$$

$$P(y_0 \mid \alpha) = P(y_0)$$

So, $P(y_2, y_1, y_0 \mid \alpha) = (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_2 - \alpha y_1 - 0)^2}{2\sigma^2}}) \cdot (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_1 - \alpha y_0 - 0)^2}{2\sigma^2}}) \cdot P(y_0)$

Take log on both sides

$$\log P(y_2, y_1, y_0 \mid \alpha) = \log \frac{1}{\sqrt{2\pi}\sigma} + -\frac{(y_2 - \alpha y_1)^2}{2\sigma^2} + \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_1 - \alpha y_0)^2}{2\sigma^2} + \log P(y_0)$$

$$\frac{\partial \log P(y_2, y_1, y_0 \mid \alpha)}{\partial \alpha} = y_1 \frac{(y_2 - \alpha y_1)}{\sigma^2} + y_0 \frac{(y_1 - \alpha y_0)}{\sigma^2}$$

$$\frac{\partial \log P(y_2, y_1, y_0 \mid \alpha)}{\partial \alpha} = \frac{(y_2 y_1 - \alpha y_1^2)}{\sigma^2} + \frac{(y_1 y_0 - \alpha y_0^2)}{\sigma^2}$$

$$0 = \frac{(y_2 y_1 - \alpha y_1^2)}{\sigma^2} + \frac{(y_1 y_0 - \alpha y_0^2)}{\sigma^2}$$

$$0 = y_2 y_1 - \alpha y_1^2 + y_1 y_0 - \alpha y_0^2$$

$$\alpha y_1^2 + \alpha y_0^2 = y_2 y_1 + y_1 y_0$$

$$\alpha (y_1^2 + y_0^2) = y_2 y_1 + y_1 y_0$$

$$\alpha = \frac{y_2 y_1 + y_1 y_0}{(y_1^2 + y_0^2)}$$

OT1

$$P(y_{n+1}, y_n, ..., y_0 \mid \alpha) = P(y_{n+1} \mid y_n, y_{n-1}, ..., y_0, \alpha) \cdot P(y_n \mid y_{n-1}, ..., y_0, \alpha) \cdot ... \cdot P(y_0 \mid \alpha)$$

Since y_{n+1} is independent from $y_{n-1}, y_{n-2}, ..., y_0$

$$P(y_{n+1}, y_n, ..., y_0 \mid \alpha) = P(y_{n+1} \mid y_n, \alpha) \cdot P(y_n \mid y_{n-1}, \alpha) \cdot ... \cdot P(y_0 \mid \alpha)$$

$$P(y_{n+1}, y_n, ..., y_0 \mid \alpha) = P(w_n = y_{n+1} - \alpha y_n) \cdot P(w_{n-1} = y_n - \alpha y_{n-1}) \cdot ... \cdot P(y_0 \mid \alpha)$$

$$P(y_{n+1}, y_n, ..., y_0 \mid \alpha) = \prod_{i=0}^{n} (P(w_i = y_{i+1} - \alpha y_i)) \cdot P(y_0 \mid \alpha)$$

$$P(y_{n+1}, y_n, ..., y_0 \mid \alpha) = \prod_{i=0}^{n} (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_{i+1} - \alpha y_i - 0)^2}{2\sigma^2}}) \cdot P(y_0)$$

Take log on both sides

$$\log P(y_{n+1}, y_n, ..., y_0 \mid \alpha) = \sum_{i=0}^{n} (\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_{i+1} - \alpha y_i - 0)^2}{2\sigma^2}) + \log P(y_0)$$

$$\frac{\partial \log P(y_{n+1}, y_n, ..., y_0 \mid \alpha)}{\partial \alpha} = \sum_{i=0}^{n} (y_i \frac{(y_{i+1} - \alpha y_i)}{\sigma^2})$$

$$\frac{\partial \log P(y_{n+1}, y_n, ..., y_0 \mid \alpha)}{\partial \alpha} = \sum_{i=0}^{n} (\frac{(y_{i+1} y_i - \alpha y_i^2)}{\sigma^2})$$

$$0 = \sum_{i=0}^{n} \left(\frac{(y_{i+1}y_i - \alpha y_i^2)}{\sigma^2}\right)$$

$$0 = \frac{\sum_{i=0}^{n} y_{i+1}y_i - \sum_{i=0}^{n} \alpha y_i^2}{\sigma^2}$$

$$0 = \sum_{i=0}^{n} y_{i+1}y_i - \sum_{i=0}^{n} \alpha y_i^2$$

$$0 = \sum_{i=0}^{n} y_{i+1}y_i - \alpha \sum_{i=0}^{n} y_i^2$$

$$\alpha \sum_{i=0}^{n} y_i^2 = \sum_{i=0}^{n} y_{i+1}y_i$$

$$\alpha = \frac{\sum_{i=0}^{n} y_{i+1}y_i}{\sum_{i=0}^{n} y_i^2}$$

T2

Using the likelihood ratio test, we got $1 = \frac{p(w_1) \cdot P(x \mid w_1)}{p(w_2) \cdot P(x \mid w_2)}$

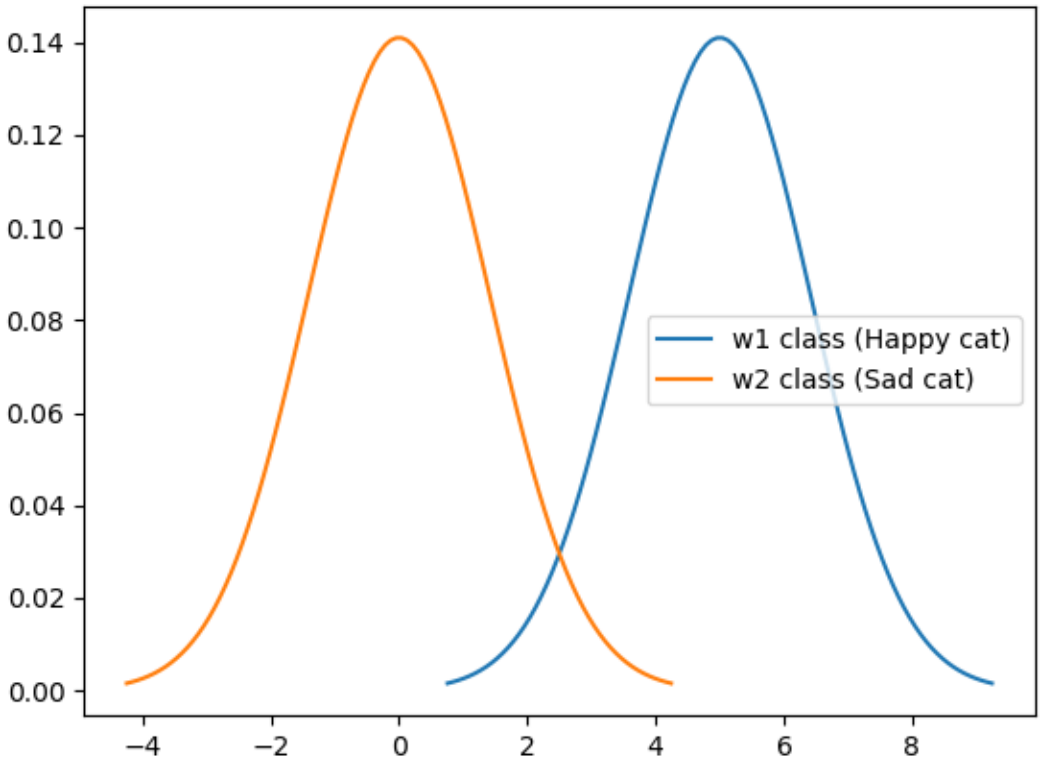
Since $p(w_1) = p(w_2)$ we can cancel both term.

Rearrange to get: $P(x \mid w_1) = P(x \mid w_2)$

$$\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2 \cdot 2^2}} = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-0)^2}{2 \cdot 2^2}}$$

By solving the equation, we got $x = 2.5$

(The posteriors plot is below)



T3

The decision boundary will shift toward the class of sad cat. Since it is more probable that the cat will be happy, the area identified as happy cat have to increase; hence, the shifting of decision boundary.

New decision boundary is found to be 1.945.

OT2

Using the likelihood ratio test, we got $1 = \frac{p(w_1) \cdot P(x \mid w_1)}{p(w_2) \cdot P(x \mid w_2)}$

Since $p(w_1) = p(w_2) = 0.5$ we can cancel both term.

Rearrange to get: $P(x \mid w_1) = P(x \mid w_2)$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$$

$$e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} = e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$$

Take log on both sides:

$$-\frac{(x-\mu_1)^2}{2\sigma^2}\log e = -\frac{(x-\mu_2)^2}{2\sigma^2}\log e$$

$$-\frac{(x-\mu_1)^2}{2\sigma^2} = -\frac{(x-\mu_2)^2}{2\sigma^2}$$

$$(x-\mu_1)^2 = (x-\mu_2)^2$$

$$x^2 - 2x\mu_1 + \mu_1^2 = x^2 - 2x\mu_2 + \mu_2^2$$

$$-2x\mu_1 + 2x\mu_2 = \mu_2^2 - \mu_1^2$$

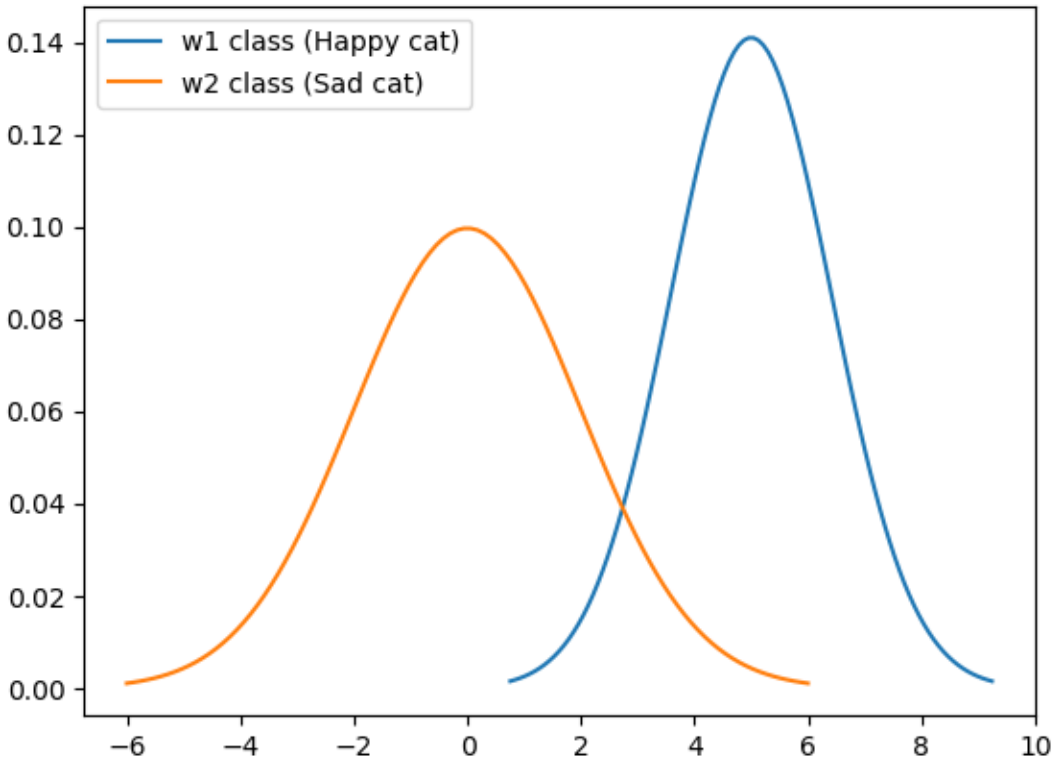
$$2x(\mu_2 - \mu_1) = (\mu_2 + \mu_1)(\mu_2 - \mu_1)$$

$$x = \frac{\mu_2 + \mu_1}{2}$$

If the student changed the distribution of sad cat to N(0,4)

The new decision boundary will be 2.7355.

(Posteriors are in the plot below)



T4

We can use Gaussian distribution to estimate some of the features. By inspecting the histogram, we can observe that Age might be the only feature that follows single Gaussian distribution. Other categorical features, such as Education and Gender cannot be used since there are a lot of bins with zero elements. Also, some of the continuous features, such as MonthlyRate, even though there are no bins with zero elements but the histogram does not follow Gaussian distribution. However, we might still be able to use GMM to estimate those features.

T5

From the histogram of all features, we can see that a lot of features that is categorical in nature, such as BusinessTravel, Department , Education etc., have a lot of bins with zero elements count. Those features cannot be mapped to any probabilistic distribution and so is not a good discretization.

T6

For age, the most sensible bin size is 10 since its histogram appears to be closest to the normal distribution.

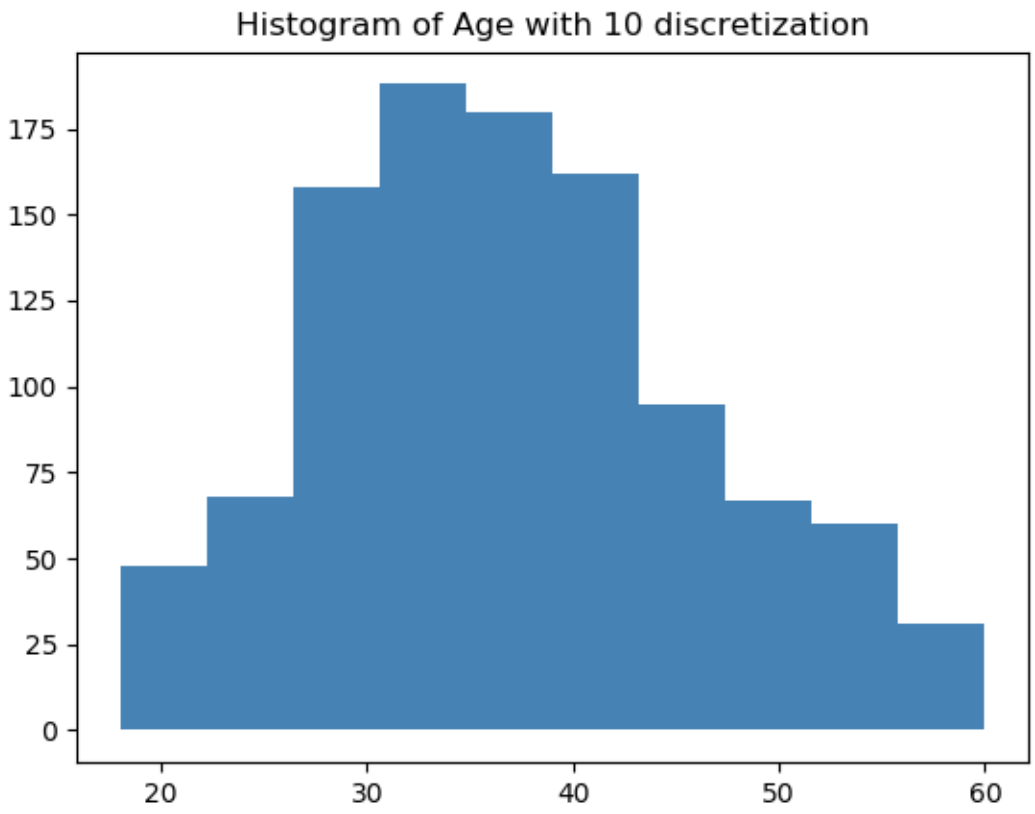
Plotting bin size more than that shows a lot of irregularity. When the bin size is 100, there are a lot of bins with zero elements since the resolution of the bin is more than the data.

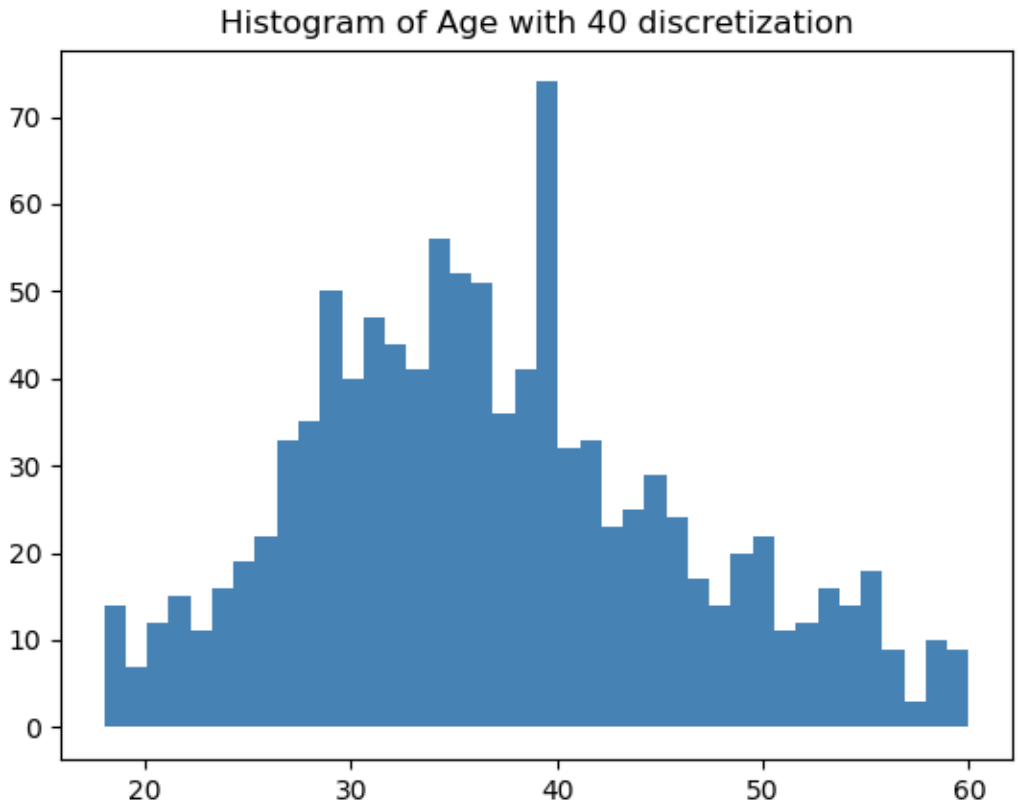
For MonthlyIncome, the most sensible bin size is 10 since its histogram appears to follow some declining pattern (decreasing at first and start to increase again near the end)

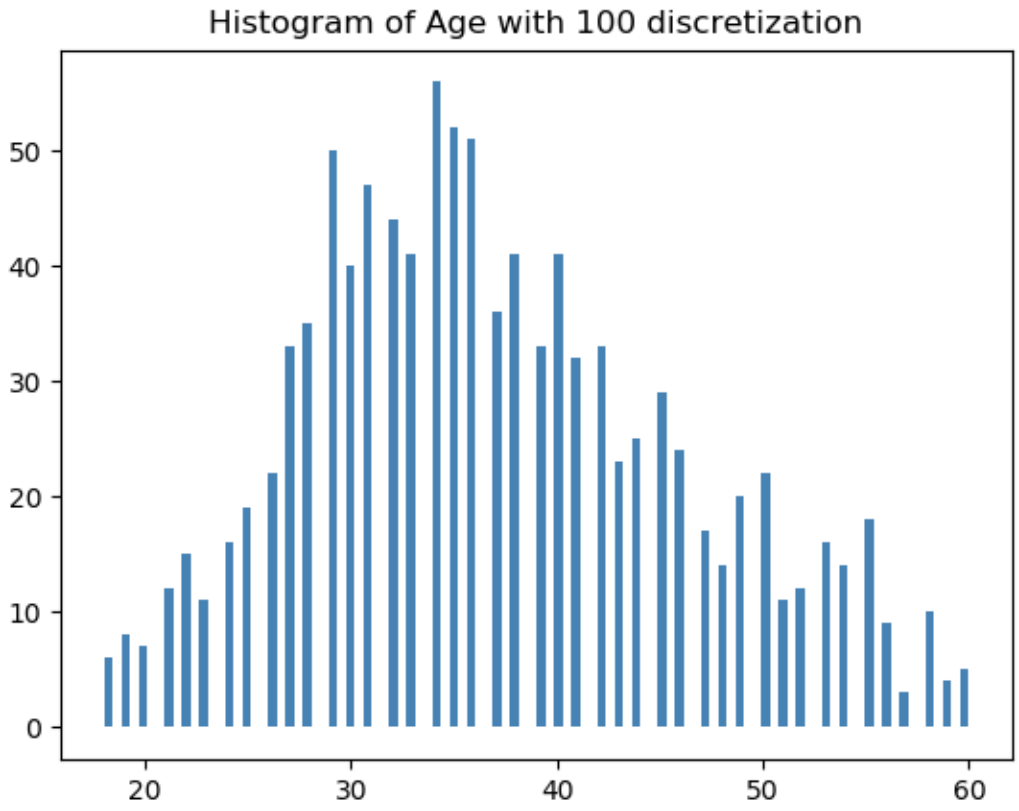
Plotting bin size more than that shows a lot of irregularity. When the bin size is 100, some of the bins have zero elements.

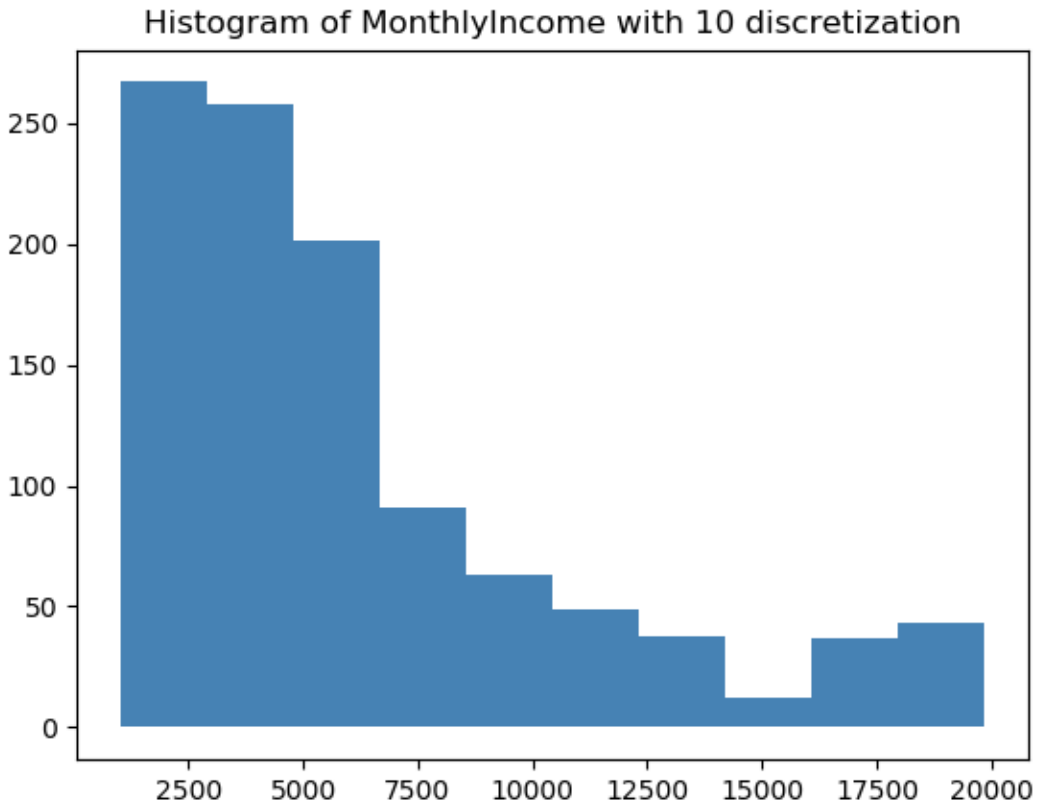
For DistanceFromHome, the most sensible bin size is 10 since it is the only bin size that does not have bin with zero elements.

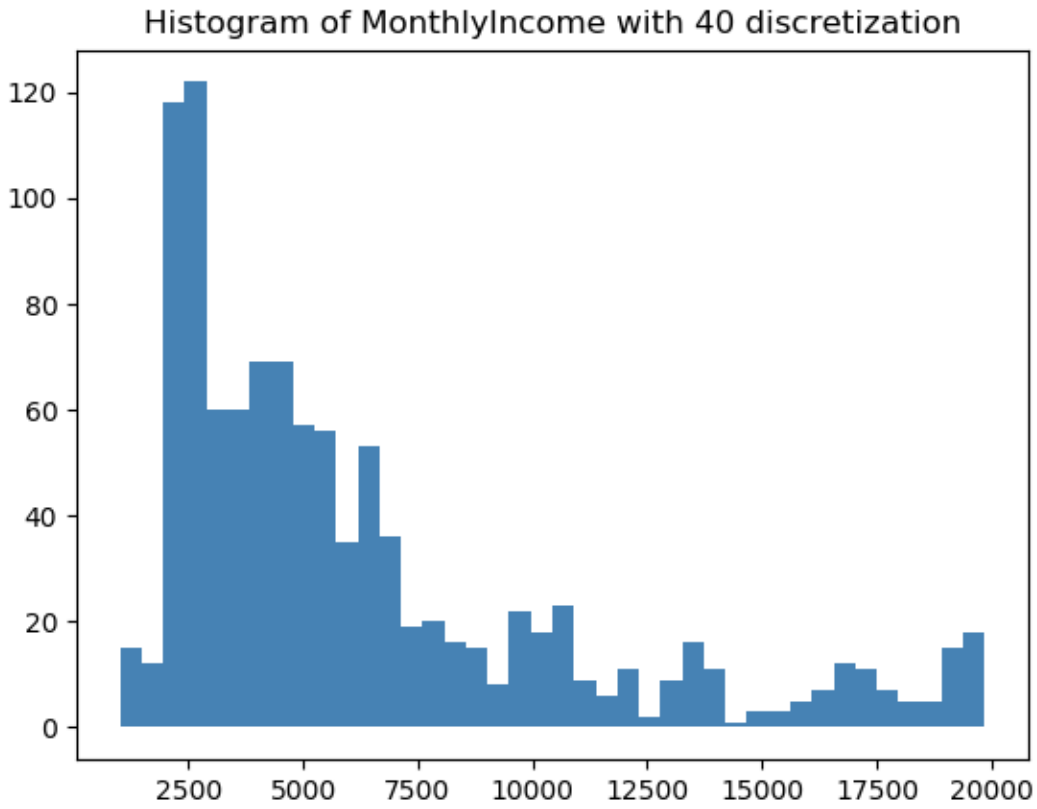
(The plots are placed below)

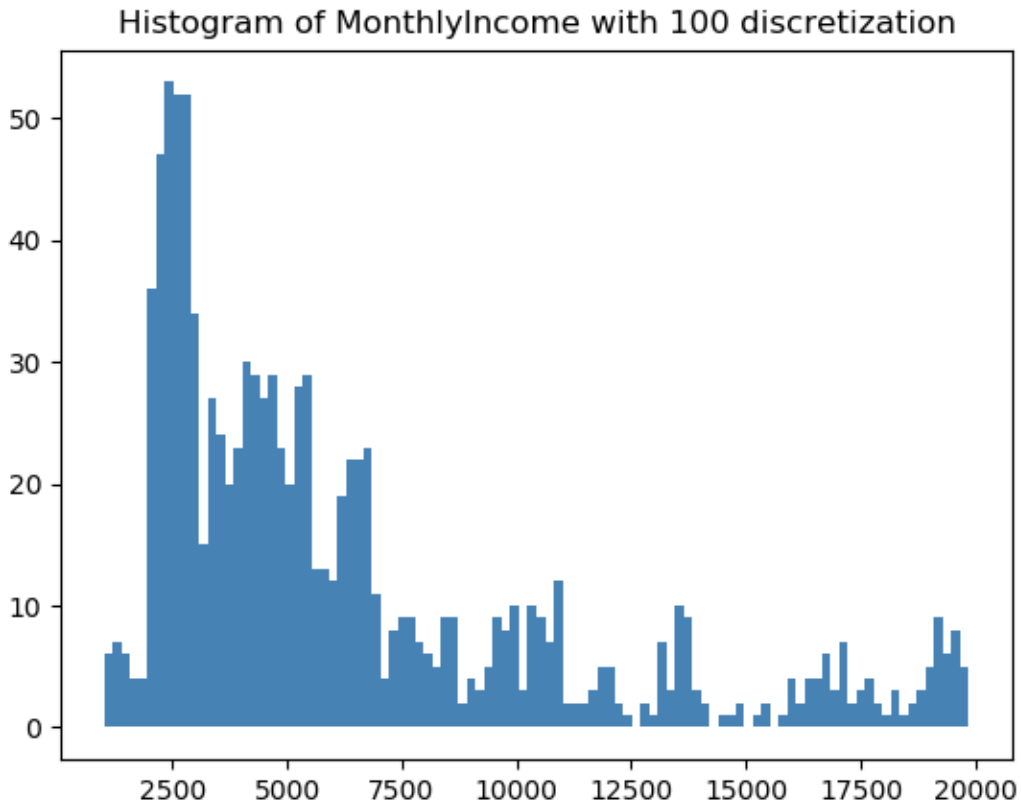


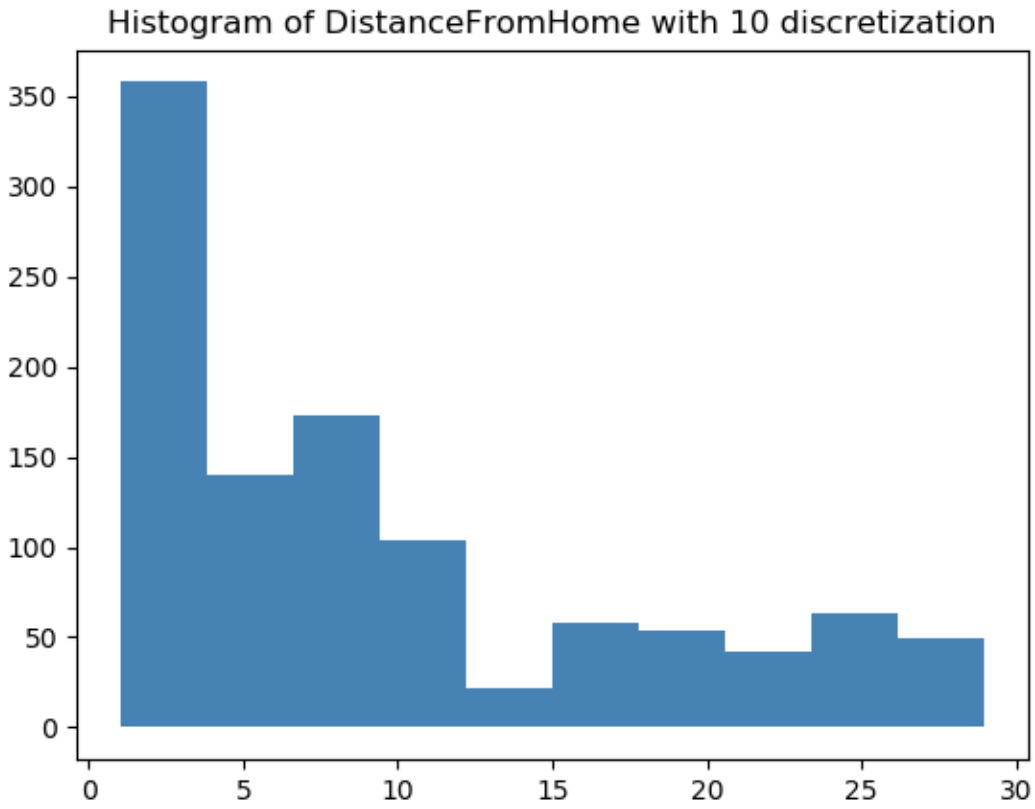

Histogram of Age with 40 discretization
70
60
50
40
30
20
10
0
20
30
40
50
60

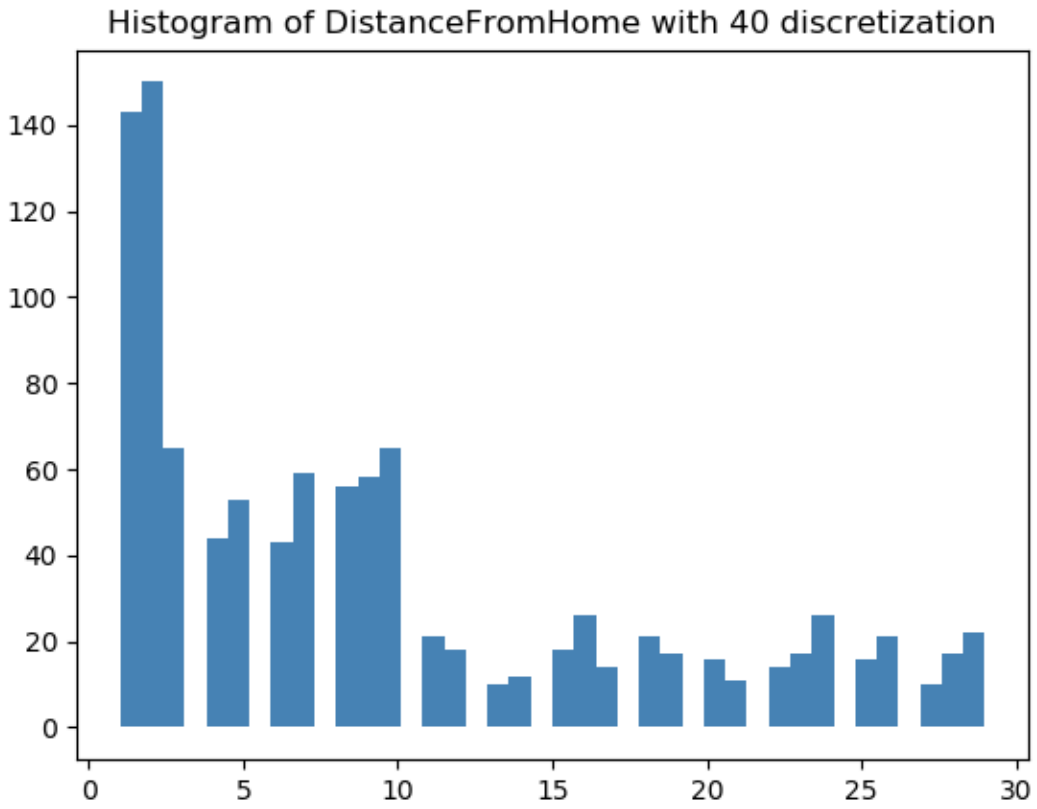

Histogram of Age with 100 discretization
50
40
30
20
10
0
20
30
40
50
60

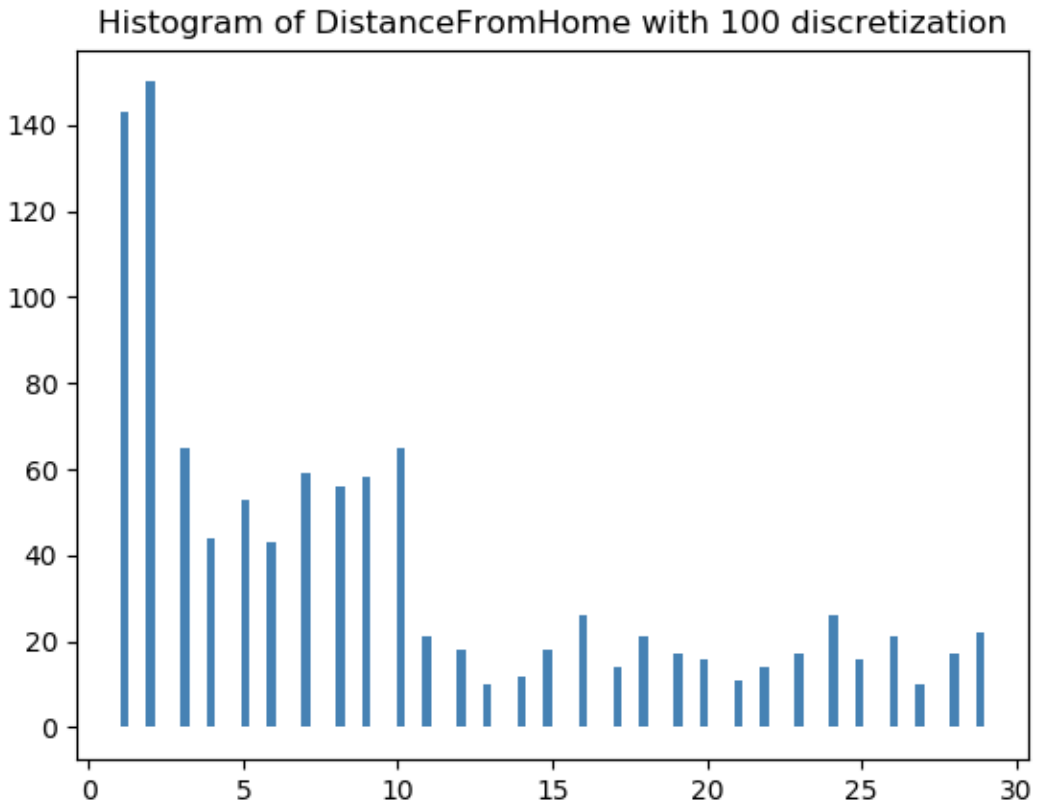

Histogram of MonthlyIncome with 10 discretization
250
200
150
100
50
0
2500
5000
7500
10000
12500
15000
17500
20000


Histogram of MonthlyIncome with 40 discretization
120
100
80
60
40
20
0
2500
5000
7500
10000
12500
15000
17500
20000


Histogram of MonthlyIncome with 100 discretization
50
40
30
20
10
0
2500
5000
7500
10000
12500
15000
17500
20000


Histogram of DistanceFromHome with 10 discretization
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30


Histogram of DistanceFromHome with 40 discretization
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30



T7

The features that are continuous in nature should be discretized.

They are Age, DailyRate, DistanceFromHome, EmployeeCount, HourlyRate, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion and YearsWithCurrManager

The features that are categorical in nature should not be discretized. They should be separated according to the number of unique categories they have.

They are BusinessTravel, Department, Education, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, Over18, OverTime, PerformanceRating, RelationshipSatisfaction, StockOptionLevel and WorkLifeBalance.

The features should be discretized if they span a large range and do not cluster into group.

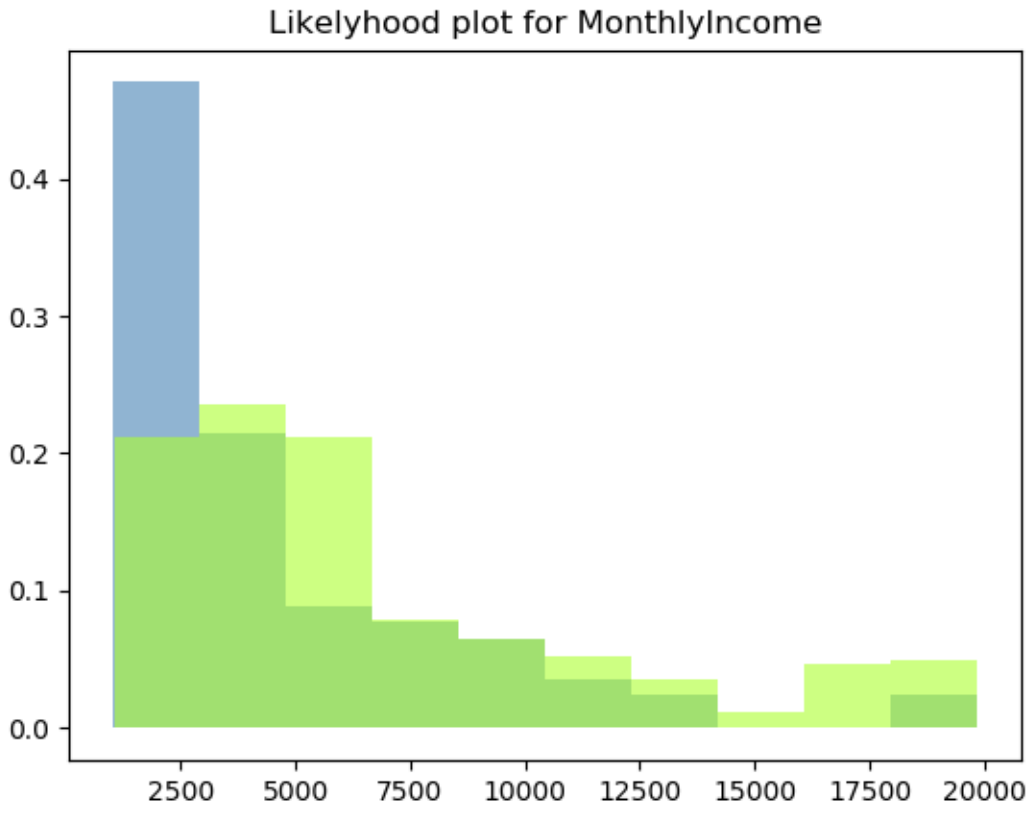
P.S. all element of StandardHours have the same value so they should not be discretized even though they are continuous.

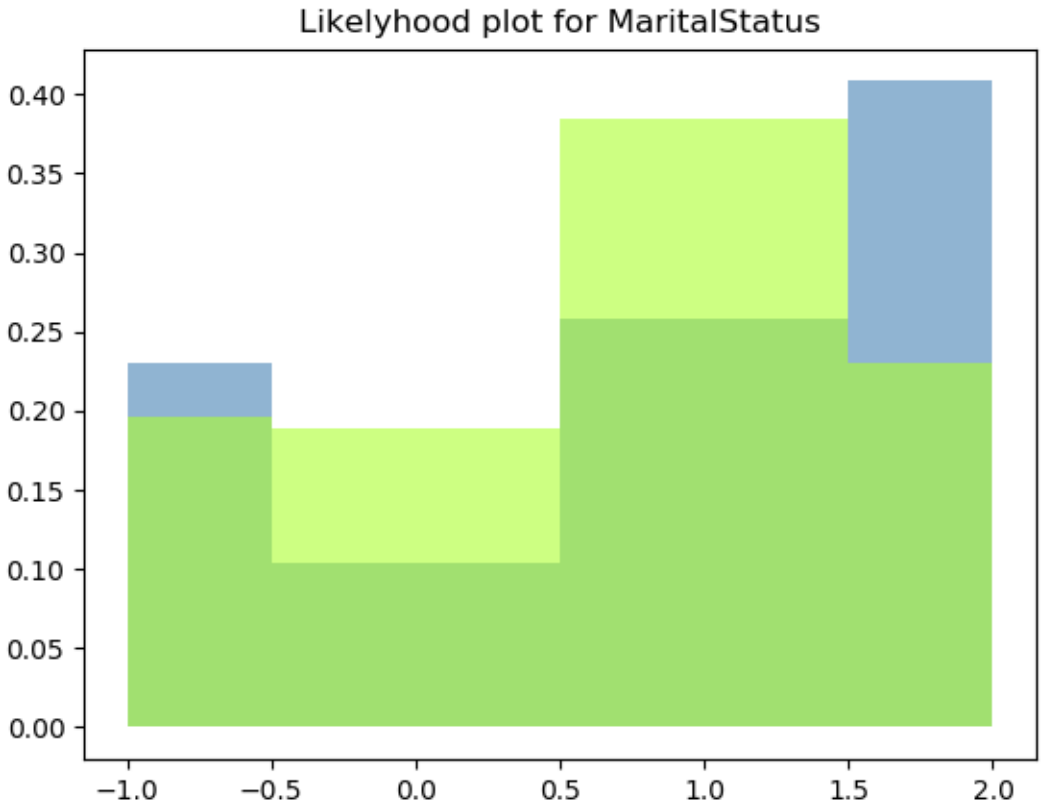
(The discretization process is in the Python code)

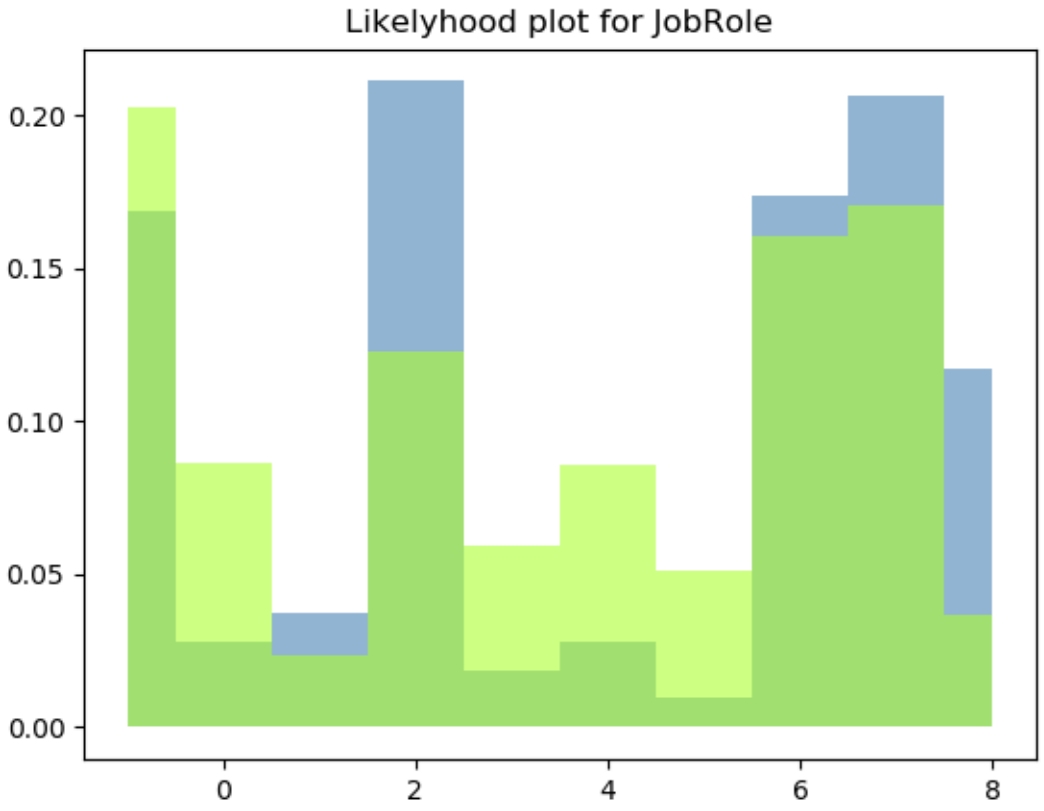
T8

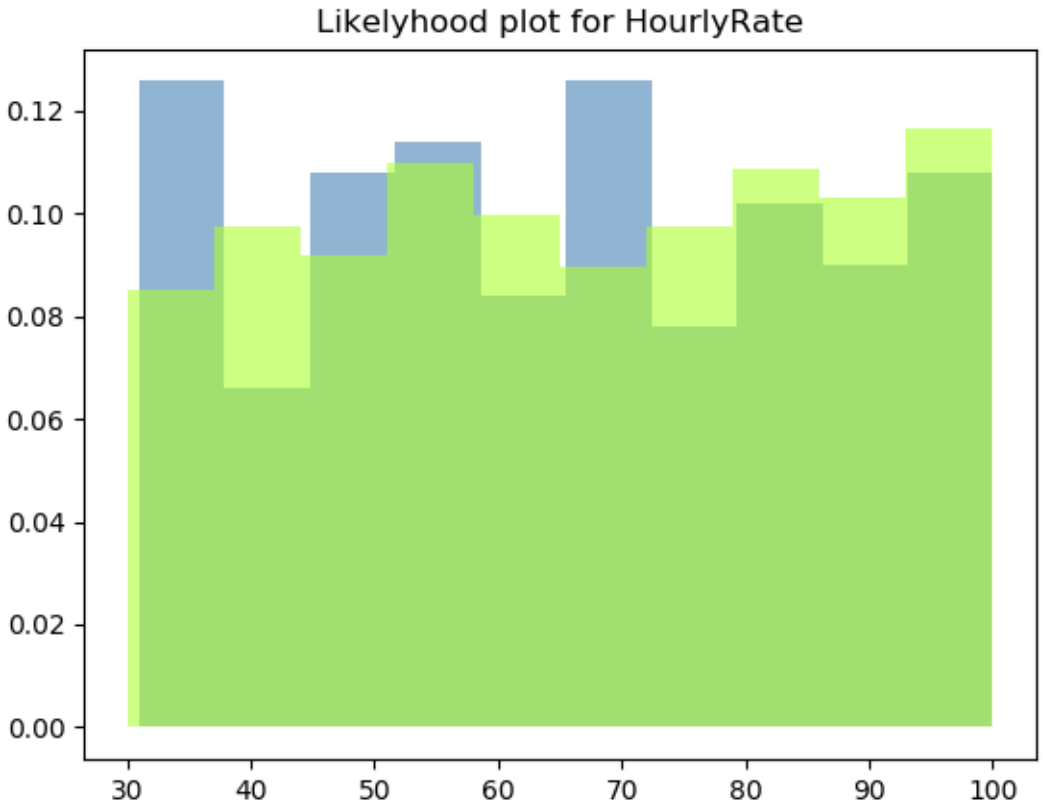
We can model the histogram with the categorical distribution. MLE of each feature can be calculated by dividing every columns in the histogram with the number of data; thus, normalizing the area under the histogram to be one. Each column in the histogram is now the estimation of maximum likelihood for the value in the range of that column.

(The plots are provided below)




Likelyhood plot for MaritalStatus
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Likelyhood plot for JobRole
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8



T9

Prior of the two classes can be calculated by dividing the number of elements belonged to that class with the number of total elements.

I got the following results when calculating data for this document.

Leave prior is 213 / 1322 = 0.16111951588502268

Stay prior is 1109 / 1322 = 0.8388804841149773

T10.

Here is the result of the classifier using histogram (categorical distribution)

Histogram model

Evaluation result:

True positive: 10

True negative: 118

False positive: 6

False negative: 14

Accuracy: 0.8648648648648649

Precision: 0.625

Recall: 0.4166666666666667

F1 score: 0.5

T11.

Here is the result of the classifier using normal distribution.

Gaussian model

Evaluation result:

True positive: 11

True negative: 107

False positive: 17

False negative: 13

Accuracy: 0.7972972972972973

Precision: 0.39285714285714285

Recall: 0.4583333333333333

F1 score: 0.4230769230769231

T12.

Here is the result of the random choice baseline classifier

Random model

Evaluation result:

True positive: 14

True negative: 65

False positive: 59

False negative: 10

Accuracy: 0.5337837837837838

Precision: 0.1917808219178082

Recall: 0.5833333333333334

F1 score: 0.288659793814433

T13.

Here is the result of the majority rule baseline classifier

Majority model

Evaluation result:

True positive: 0

True negative: 124

False positive: 0

False negative: 24

Accuracy: 0.8378378378378378

Precision: -1

Recall: 0.0

F1 score: 0.0

T14

From the comparison, we found out that both of the models we created perform better than the random baseline model in almost every aspect.

Histogram model accuracy 0.865 > 0.533 of random model

Histogram model precision 0.625 > 0.191 of random model

Histogram model recall 0.417 < 0.583 of random model

Histogram model F1 0.5 > 0.289 of random model

Gaussian model accuracy 0.797 > 0.533 of random model

Gaussian model precision 0.393 > 0.191 of random model

Gaussian model recall 0.458 < 0.583 of random model

Gaussian model F1 0.423 > 0.289 of random model

For the recall, since the label of the data is not uniformly distributed. By making random choice, it is more possible that the random model will be picking Attrition=1 more than the model we created.

When comparing with the majority model

Histogram model accuracy 0.865 > 0.838 of majority model

Histogram model recall 0.417 > 0 of majority model

Histogram model F1 0.5 > 0 of majority model

Gaussian model accuracy 0.797 < 0.838 of majority model

Gaussian model recall 0.458 > 0 of majority model

Gaussian model F1 0.423 > 0 of majority model

The majority model do not have precision value since it never predicts 1, and the recall =0 because it never answers 0. This makes it impossible to calculate sensible F1 score. As a result, I will use only accuracy to compare the model.

From the comparison, we can see that our histogram model has higher accuracy than the majority model but the Gaussian one has lower accuracy.

However, it is because our data is not uniformly distributed. There is about 83 percent of answer 0 so it is sensible that the majority model has high accuracy.

Nevertheless, our histogram model still performs better than the majority model in every aspect.

T15

By running both histogram version and normal distribution of the model against various thresholds, we obtained the following results:

Histogram model

Max accuracy is 0.8918918918918919 occurred at threshold = 2.40

Max precision is 1.0 occurred at threshold = 2.40

Max recall is 0.8333333333333334 occurred at threshold = -5.0

Max f1 score is 0.5263157894736842 occurred at threshold = -1.60

Gaussian model

Max accuracy is 0.8851351351351351 occurred at threshold = 0.85

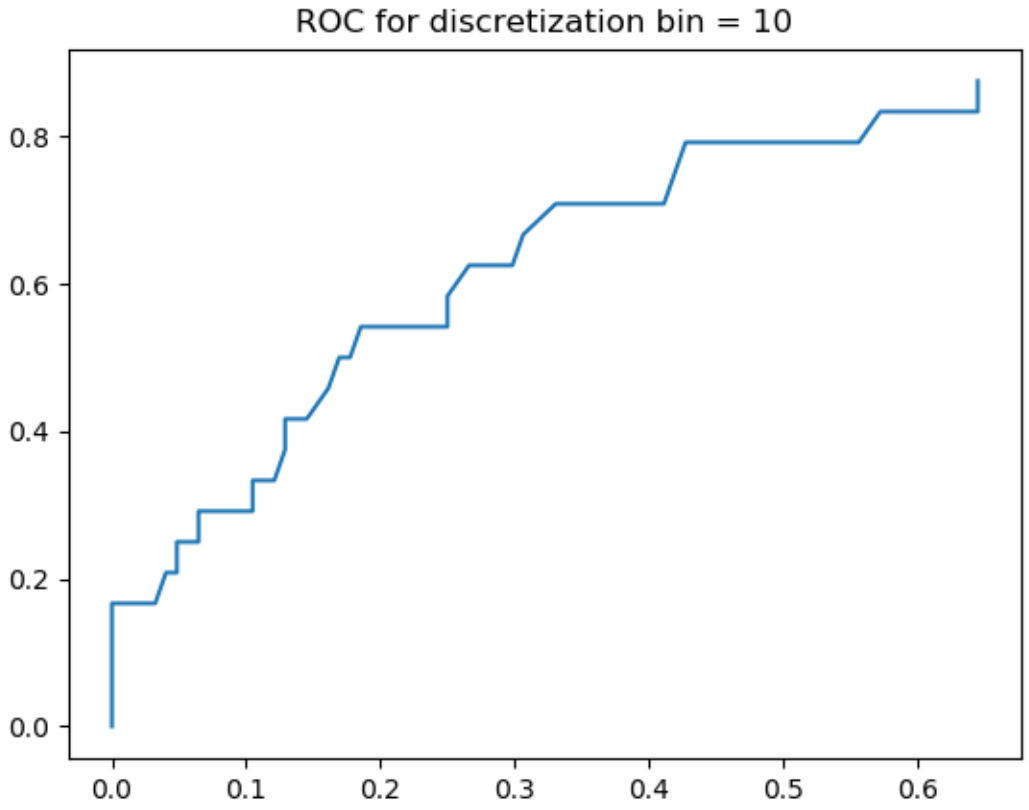
Max precision is 1.0 occurred at threshold = 1.30

Max recall is 0.9583333333333334 occurred at threshold = -5.0

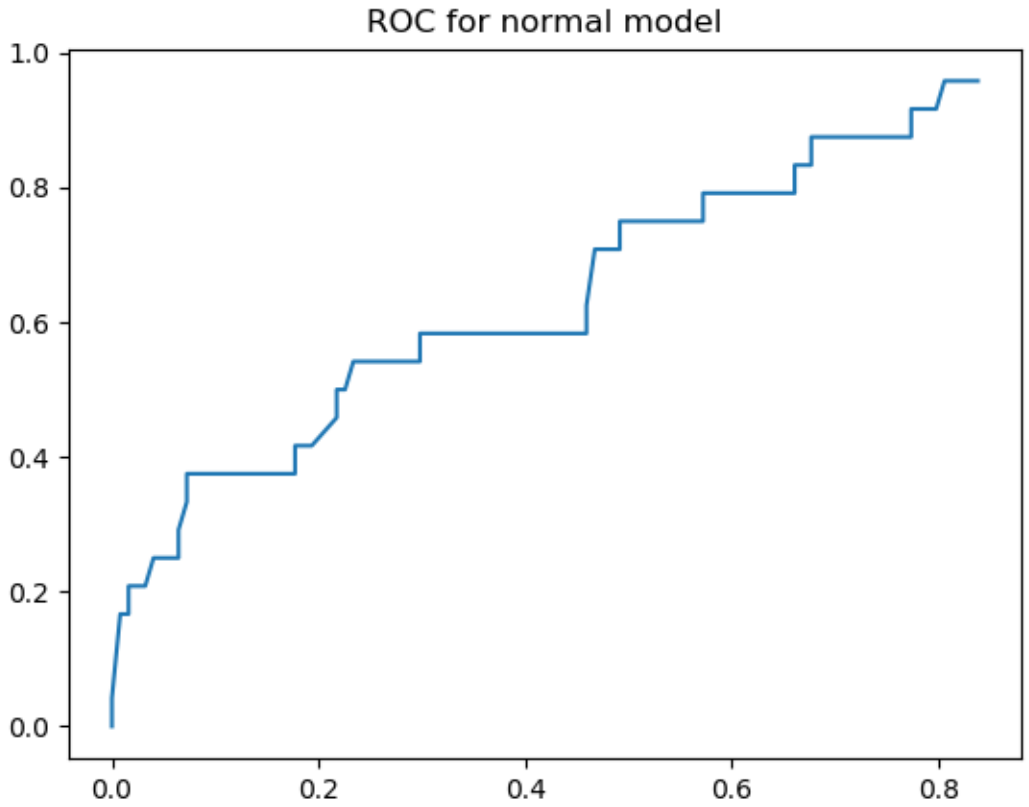
Max f1 score is 0.5142857142857142 occurred at threshold = 0.85

T16

Histogram model



Gaussian model



T17

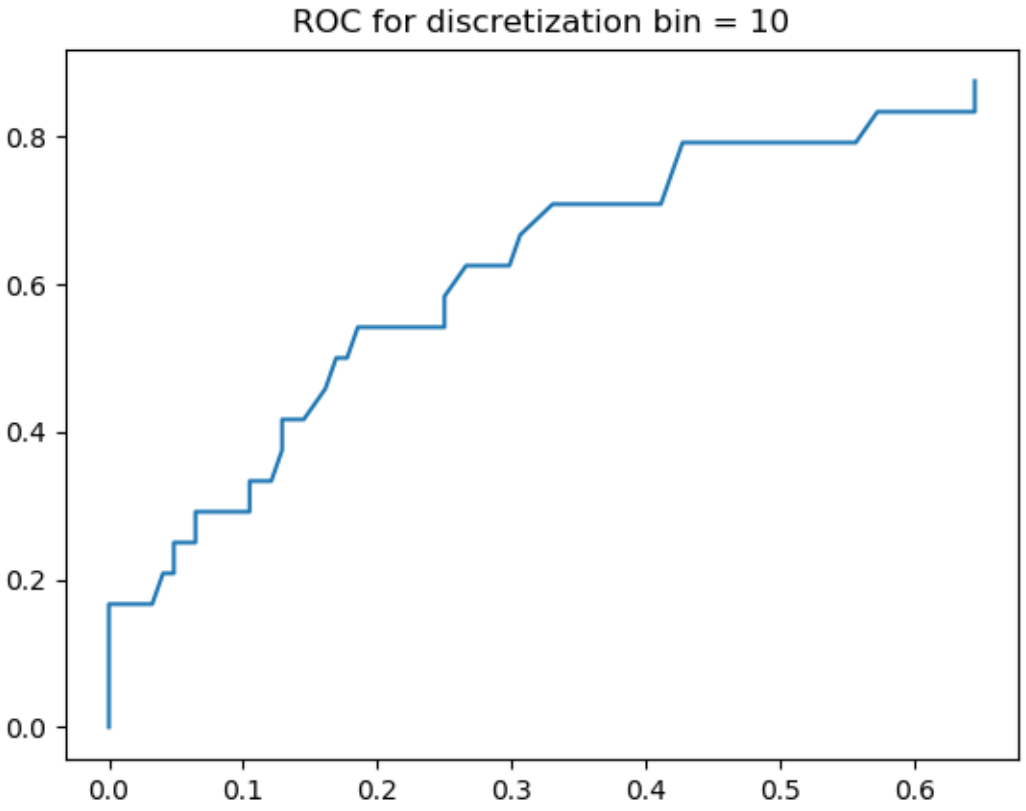
When changing the discretization bin to 5, running the histogram model against various thresholds yield the following result.

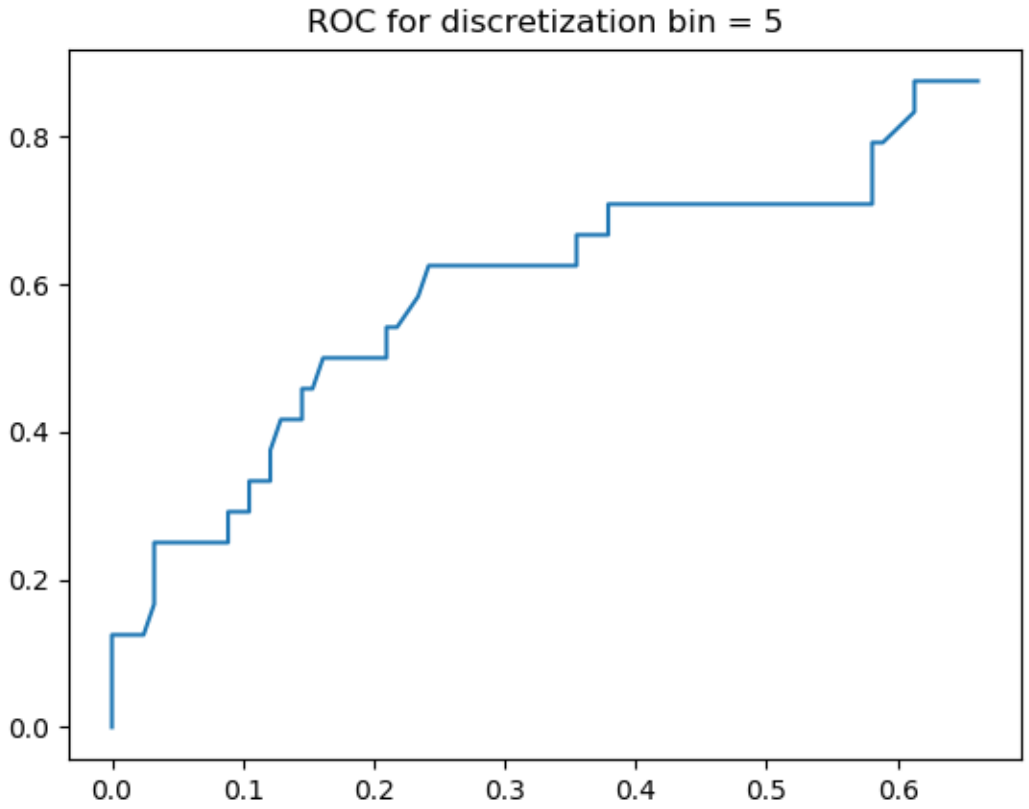
Max accuracy is 0.8783783783783784 occurred at threshold = 1.35

Max precision is 1.0 occurred at threshold = 1.95

Max recall is 0.875 occurred at threshold = -5.0

Max f1 score is 0.5531914893617021 occurred at threshold = -0.30


ROC for discretization bin = 10
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6



From the observation, we found that discretization bin=10 is better than bin=5 in my case.

Looking at the graph, we can observe that as the false positive rate increase, the recall rate of bin=10 is increasing at a greater rate than that of bin=5.

As a result, the area under the curve of bin=10 is greater than that of bin=5.

This allows us to conclude that the discretization bin=10 model performs slightly better than bin=5.

OT3

The mean accuracy of the model is 0.8405405405405407

The variant of the model is 0.00033966398831263727