Mini Project 1: Naïve – Bayes Classification

1. Instructions on compiling and running the programs

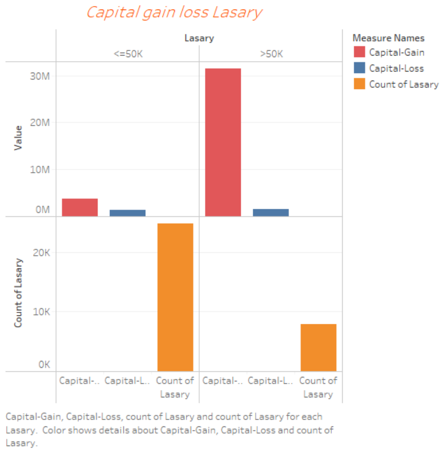
data	Data_train input (adults)
data_simulate	Data after simulation
data_used	Data to use for modeling
data_set_1_train	Data_train with salary <=50K
data_set_2_train	Data_train with salary >50K
Function	
find_prob_discrete_distribution	Find out the probability density function for discrete variablce
simulation_runif_acording_probilaty	Simulation: input: columns of data
simulation_runif_acording_probilaty_0	Simulation: input: probability
simulation_normal_distribution_discrete	Gaussian distribution
bayer_discrete(vector,data)	Bayes for discrete
bayer_gauus(vector,data)	Bayes for gaussian
bayer_classified(data_test,data_train)	Model bayel classified
bining(data,list_columns,list_bin_group)	Bining with columns data and number of groups
Kfold(data)	Kfold with data train 90%, test 10%

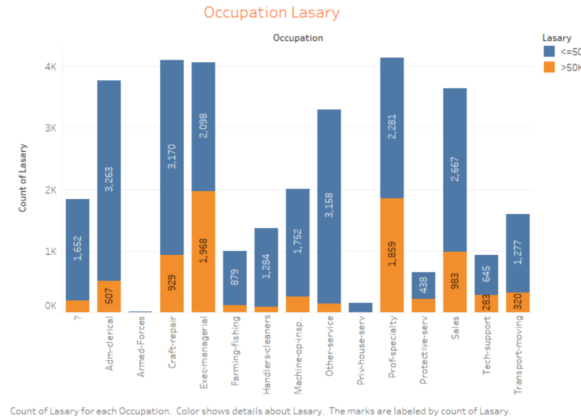
2. Description of the main steps

a. Explanatory Data analysis

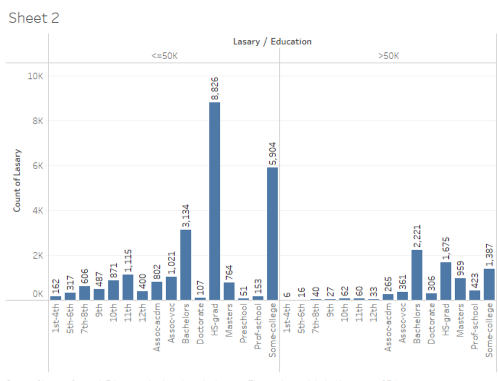
This dataset is imbalanced since approximately 75% of the dataset included the people with salary less than or equal 50k and 25% of the dataset include the people with salary more than 50k.

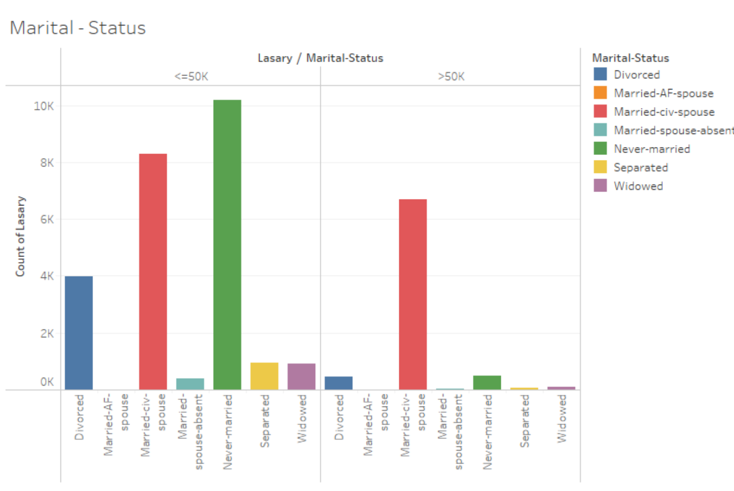
The chart below shows the relationship between the capital gain with the income salary. The capital gain for people who have the income salary greater than 50k is much larger than those for people who have the income salary less than or equal to 50k, while the capital loss for people who have the income salary greater than 50k and people with the income salary less than or equal to 50k is approximately the same. Additionally, 100% of people work as a private house service would receive a salary less than or equal to 50k. Furthermore, people who work as a farming fishing, Handlers-cleaners, Machine-op-inspector, and other services tends to receive a salary less than or equal to 50k, when the proportion of people with salary less than or equal to 50k dominated those with salary greater than 50k.





The bar chart below illustrates the relationship between the salary and the education. From the bar graph, it is hard to conclude whether the level of education affects the salary or not. In contrast, from the map graph, almost every people who does not live in the US tend to get the salary less than or equal to 50k.





There is a huge difference of the salary with the people whose marital status is "single" (both divorced and never married) when they will to receive the money less than or equal to 50k.

b. Handling Missing Value

2 strategies to handing missing value:

- Remove all the missing value: Since the total missing value in this data is just about 7% of the data, therefore, we can easily remove all the missing data without affecting the performance of the model.
- Importing all the missing value: The strategies to import the value is to calculate the proportion of other value and then randomly selected the new value for the missing value based on the proportion.

c. Oversampling

As I mentioned from the first part of this documentation, since the data is imbalanced, therefore, if we still use the data to for modeling, the result can be good, but it does not have any meaningful statistics. To solve this problem, oversampling, which is make the data more balance to use, should be used in this dataset.

Main idea for oversampling the dataset:

For discrete variable: Find out the probability of each value with the salary greater than 50k. Then impute the value for the data at random based on that probability so that the probability can still be the same.

For continuous variable:

- Use binning-width method to transform all the attribute into a categorial attribute. Then use the same concept with discrete variable
- Gaussian distribution: Assume the continuous variable as Normal distribution with mean and standard deviation. Find out the probability for each of value in based on the cumulative density function.

3. Evaluation

I. Oversampling

a. Use Naïve – Bayes Classifier to test in testing data

Data1: Impute missing value and use binning width for continuous variable

	C1	C2
C1	8468	2892
C2	501	3199

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7747

P(precision) = t-pos / (t-pos + f-pos) = 0.7454

R(Recall) = t-pos / (t-pos + f-neg) = 0.9441

F-1 measure = 2*(P*R)/(P+R) = 0.8331

Data2: Impute missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8482	2878
C2	489	3211

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7764
P(precision) = t-pos / (t-pos + f-pos) = 0.7467
R(Recall) = t-pos / (t-pos + f-neg) = 0.9455
F-1 measure = 2*(P*R)/(P+R) = 0.8344

Data3: Remove missing value and use binning width for continuous variable

	C1	C2
C1	8511	2849
C2	480	3220

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7789
P(precision) = t-pos / (t-pos + f-pos) = 0.7492
R(Recall) = t-pos / (t-pos + f-neg) = 0.9466
F-1 measure = 2*(P*R)/(P+R) = 0.8364

Data4: Remove missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8445	2915
C2	493	3207

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7737
P(precision) = t-pos / (t-pos + f-pos) = 0.7434
R(Recall) = t-pos / (t-pos + f-neg) = 0.9448
F-1 measure = 2*(P*R)/(P+R) = 0.8321

b. K-fold cross validation

Data1: Impute missing value and use binning width for continuous variable

	C1	C2
C1	8568	2792
C2	482	3218

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7826
P(precision) = t-pos / (t-pos + f-pos) = 0.7542
R(Recall) = t-pos / (t-pos + f-neg) = 0.9467
F-1 measure = 2*(P*R)/(P+R) = 0.8396

Data2: Impute missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8571	2795
C2	488	3212

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7821
P(precision) = t-pos / (t-pos + f-pos) = 0.7541
R(Recall) = t-pos / (t-pos + f-neg) = 0.9461
F-1 measure = 2*(P*R)/(P+R) = 0.8393

Data3: Remove missing value and use binning width for continuous variable

	C1	C2
C1	8435	2925
C2	520	3180

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7712
P(precision) = t-pos / (t-pos + f-pos) = 0.7425
R(Recall) = t-pos / (t-pos + f-neg) = 0.9419
F-1 measure = 2*(P*R)/(P+R) = 0.8304

Data4: Remove missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8670	2690
C2	525	3175

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7866
P(precision) = t-pos / (t-pos + f-pos) = 0.7632
R(Recall) = t-pos / (t-pos + f-neg) = 0.9429
F-1 measure = 2*(P*R)/(P+R) = 0.8436

II. Under sampling

a. Use Naïve – Bayes Classifier to test in testing data

Data1: Impute missing value and use binning width for continuous variable

	C1	C2
C1	8792	2586
C2	1590	2110

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.723
P(precision) = t-pos / (t-pos + f-pos) = 0.7727
R(Recall) = t-pos / (t-pos + f-neg) = 0.8469
F-1 measure = 2*(P*R)/(P+R) = 0.8081

Data2: Impute missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8096	3264
C2	2021	1679

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.6491
P(precision) = t-pos / (t-pos + f-pos) = 0.7127
R(Recall) = t-pos / (t-pos + f-neg) = 0.8002
F-1 measure = 2*(P*R)/(P+R) = 0.7539

Data3: Remove missing value and use binning width for continuous variable

	C1	C2
C1	7967	3393
C2	1659	2041

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.6645
P(precision) = t-pos / (t-pos + f-pos) = 0.7013
R(Recall) = t-pos / (t-pos + f-neg) = 0.8277
F-1 measure = 2*(P*R)/(P+R) = 0.7593

Data4: Remove missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8068	3292
C2	1615	2085

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.6742
P(precision) = t-pos / (t-pos + f-pos) = 0.7102
R(Recall) = t-pos / (t-pos + f-neg) = 0.8332
F-1 measure = 2*(P*R)/(P+R) = 0.7668

b. K-fold cross validation

Data1: Impute missing value and use binning width for continuous variable

	C1	C2
C1	8789	2571
C2	1585	2115

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.724
P(precision) = t-pos / (t-pos + f-pos) = 0.806
R(Recall) = t-pos / (t-pos + f-neg) = 0.8472
F-1 measure = 2*(P*R)/(P+R) = 0.8261

Data2: Impute missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8792	2568
C2	1590	2110

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7239
P(precision) = t-pos / (t-pos + f-pos) = 0.7739
R(Recall) = t-pos / (t-pos + f-neg) = 0.8469
F-1 measure = 2*(P*R)/(P+R) = 0.8088

Data3: Remove missing value and use binning width for continuous variable

	C1	C2
C1	8815	2545
C2	1580	2120

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7261
P(precision) = t-pos / (t-pos + f-pos) = 0.776
R(Recall) = t-pos / (t-pos + f-neg) = 0.848
F-1 measure = 2*(P*R)/(P+R) = 0.8104

Data4: Remove missing value and use Gaussian distribution for continuous variable

	C1	C2
C1	8860	2500
C2	1578	2122

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) = 0.7292
P(precision) = t-pos / (t-pos + f-pos) = 0.7799
R(Recall) = t-pos / (t-pos + f-neg) = 0.8488
F-1 measure = 2*(P*R)/(P+R) = 0.8115

Based on the result from above, oversampling shows better performance than under sampling. Furthermore, remove missing value and use Gaussian distribution for continuous variable seems to show the best performance among all of these handling missing value and handling continuous variable.