Prerequisite

All code and figures are contained in *https://github.com/feedlord18/CSDS313.git*

Problem 1

From the results are can see none of the determining variables are selected. The top index and respective coefficients are:

Index	p-Value
608	0.9994125
585	0.9990805
64	0.9935824
621	0.9934949
724	0.9931771
595	0.9931551
920	0.9924055
222	0.9910656
743	0.9906898
446	0.9901487

Table 1: Top 10 Variables Ranked by Pearson Correlation

From the model summary, we can see no significant variables. This aligns with our previous conclusion, which none of the determining variables are within the 10 variables. Therefore none of the 10 variables contribute to the outcome.

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.2685 -2.2622 -0.4485  1.6858  8.7561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1647295  0.3286595  -0.501    0.617
V1          -0.0011903  0.3396155  -0.004    0.997
V2           0.0003819  0.3062771   0.001    0.999
V3          -0.0037945  0.3207404  -0.012    0.991
V4           0.0025225  0.3122940   0.008    0.994
V5           0.0014030  0.3305957   0.004    0.997
V6          -0.0041425  0.3220144  -0.013    0.990
V7          -0.0020092  0.3045516  -0.007    0.995
V8          -0.0063696  0.3701561  -0.017    0.986
V9           0.0030180  0.3022472   0.010    0.992
V10         -0.0057993  0.3267589  -0.018    0.986

Residual standard error: 3.106 on 89 degrees of freedom
Multiple R-squared:  1.029e-05, Adjusted R-squared:  -0.1123
F-statistic: 9.155e-05 on 10 and 89 DF,  p-value: 1
```

We perform lasso regression with $\alpha = 1$ and $\lambda = 0.1, 1, 10$. All variables except V1 (608)

is dropped. Still again, none of the 5 initial determining variables are in the 10 variables selected, so it is expected.

```
> coef(model_0.1)
11 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -0.1666942
V1           0.0000000
V2           .
V3           .
V4           .
V5           .
V6           .
V7           .
V8           .
V9           .
V10          .
```

```
> coef(model_1)
11 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -0.1666942
V1           0.0000000
V2           .
V3           .
V4           .
V5           .
V6           .
V7           .
V8           .
V9           .
V10          .
```

```
> coef(model_10)
11 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -0.1666942
V1           0.0000000
V2           .
V3           .
V4           .
V5           .
V6           .
V7           .
V8           .
V9           .
V10          .
```

We apply discretization on all feature values and below are top 10 variable indexes ranked according to their concordance scores. We see out of the 10 features, 33, 902, 768, 261 are within the list, which are initial determining variables to the outcome. The determining values are highlighted in bold.

Index	concordance-Scores
33	0.71
902	0.67
768	0.64
8	0.63
18	0.63
175	0.63
261	0.63
799	0.63
53	0.62
202	0.62

Table 2: Top 10 Variables Ranked by Concordance Scores

Below, we create the naive Bayesian conditional probability table. We only need to record for one binary variable because 0/1 are complementary. Therefore with a binary feature, we only need one variable to record the CPT.

	Y = 0	Y = 1
$x_{33} = 1$	15	33
$x_{902} = 1$	19	33
$x_{768} = 1$	17	28
$x_8 = 1$	21	31
$x_{18} = 1$	20	30
$x_{175} = 1$	19	29
$x_{261} = 1$	20	30
$x_{799} = 1$	21	31
$x_{53} = 1$	16	25
$x_{202} = 1$	22	31

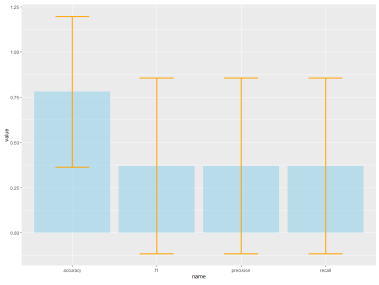
Table 3: NB CPT

Using LOOCV (Leave One Out Cross Validation) we can train our naive Bayesian model on each 99 samples and validate our results on the dropout sample. Therefore we can achieve the following metrics (This is results calculated after all 100 validation steps are finished, I couldn't understand how we would obtain variance because precision and recall and f1 would always be equal in a single validation test):

metrics	value
accuracy	0.78
precision	0.755
recall	0.787
f1	0.771

Table 4: Metrics of 100 Validation Prediction Results

This is the results from average single validation metrics, the results are vastly different from the previous result for precision, recall, and f1:



metrics	value	variance
accuracy	0.78	0.173
precision	0.37	0.235
recall	0.37	0.235
f1	0.37	0.235

We generate 50 test samples from the same mechanism as the training data. Then generate the CPT from all 100 samples of training data and evaluate the model with the test data. From below we can see the ROC curve and Precision-Recall curve which visualizes the performance of the model. The ROC curve demonstrates an AUC larger than 0.5, which suggests it performs better than a random classifier.