Classification

Marcela Pantoja

How Linear Models for Classification work: Linear models for classification work by creating decision boundaries in the data to separate the observations into different regions, where most of the observations in a region are of the same class. The strengths of these liner models is in their simplicity, but since there weakness is that they may under fit the data due to its tendency to try and find a linear decision boundary.

Reading in a CSV file about Business Price Index's in September to perform Logistic Regression on. Source: https://stats.govt.nz/large-datasets/csv-files-for-download/ (https://stats.govt.nz/large-datasets/csv-files-for-download/)

```
df <- read.csv("Buss_Price_Index_Sept_2022.csv", header=TRUE)
```

Data Cleaning and turning Qualatative values into Factors

```
df <- df[,c(1,2,3,6,7,8)]

df$Subject <- factor(df$Subject)
df$Group <- factor(df$Group)
df$Series_title_1 <- factor(df$Series_title_1)
```

Filling in any NA quantatative values with the median

```
df$Data_value[is.na(df$Data_value)] <- median(df$Data_value, na.rm=T)
df$Period[is.na(df$Period)] <- median(df$Period, na.rm=T)
```

Setting the seed and dividing the data set into an 80/20 train test

```
set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*0.80, replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

R Functions for Data Exploration

```
head(df) #first 6 rows
```

	Series_reference <chr>	Period <dbl>	Data_value <dbl>	Subject <fct>
1	CEPQ.S2371	1996.12	899	Capital Goods Price Index - CEP
2	CEPQ.S2371	1997.03	884	Capital Goods Price Index - CEP
3	CEPQ.S2371	1997.06	925	Capital Goods Price Index - CEP
4	CEPQ.S2371	1997.09	932	Capital Goods Price Index - CEP
5	CEPQ.S2371	1997.12	929	Capital Goods Price Index - CEP
6	CEPQ.S2371	1998.03	940	Capital Goods Price Index - CEP

6 rows | 1-5 of 7 columns

```
tail(df) #Last 6 rows
```

	Series_reference	Period	Data_value	Subject	▸
	<chr>	<dbl>	<dbl>	<fct>	
74301	PPIQ.SQURS2110	2021.06	1303	Producers Price Index - PPI	
74302	PPIQ.SQURS2110	2021.09	1328	Producers Price Index - PPI	
74303	PPIQ.SQURS2110	2021.12	1342	Producers Price Index - PPI	
74304	PPIQ.SQURS2110	2022.03	1391	Producers Price Index - PPI	
74305	PPIQ.SQURS2110	2022.06	1432	Producers Price Index - PPI	
74306	PPIQ.SQURS2110	2022.09	1459	Producers Price Index - PPI	

6 rows | 1-5 of 7 columns

```
names(df) #column names
```

```
## [1] "Series_reference" "Period"           "Data_value"       "Subject"
## [5] "Group"            "Series_title_1"
```

```
dim(df) # column row dimesnions
```

```
## [1] 74306     6
```

```
summary(df) #Summary stats for each column
```

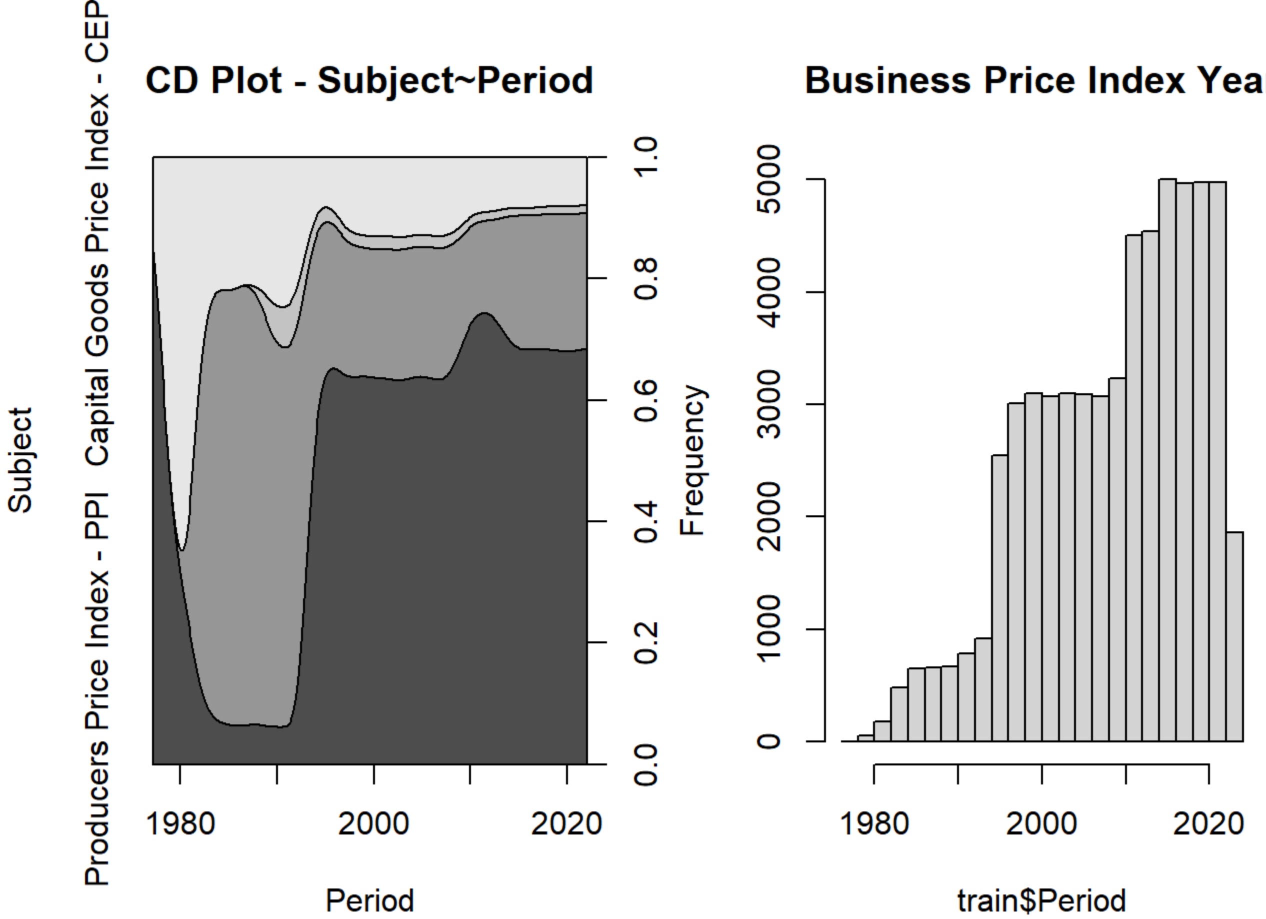
```
##  Series_reference       Period       Data_value  
##  Length:74306       Min.   :1977   Min.   :  -5.8  
##  Class :character   1st Qu.:2001   1st Qu.: 780.0  
##  Mode  :character   Median :2010   Median : 998.0  
##                     Mean   :2008   Mean   : 964.6  
##                     3rd Qu.:2016   3rd Qu.:1105.0  
##                     Max.   :2022   Max.   :5239.0  
##                                                    
##                              Subject     
##  Capital Goods Price Index - CEP: 8255  
##  Energy Statistics - NRG        : 1329  
##  Farm Inputs - FPI              :17952  
##  Producers Price Index - PPI    :46770  
##                                         
##                                         
##                                         
##                                                                       Group      
##  Farm expense price index - Expense categories -  (Base Dec 2013 = 1000):17952  
##  Inputs (ANZSIC06) - NZSIOC level 4, Base: Dec. 2010 quarter (=1000)    : 6228  
##  Inputs (ANZSIC06) - NZSIOC level 3, Base: Dec. 2010 quarter (=1000)    : 6226  
##  Outputs (ANZSIC06) - NZSIOC level 4, Base: Dec. 2010 quarter (=1000)   : 5928  
##  Outputs (ANZSIC06) - NZSIOC level 3, Base: Dec. 2010 quarter (=1000)   : 5814  
##  Published output commodities, Base Dec 2009                            : 5790  
##  (Other)                                                                :26368  
##                                                      Series_title_1 
##  All Industries                                            : 1070  
##  Forestry and Logging                                      :  684  
##  Mining                                                    :  684  
##  Owner-Occupied Property Operation (National Accounts Only):  684  
##  Printing                                                  :  684  
##  Wholesale Trade                                           :  684  
##  (Other)                                                   :69816  
```

```
str(df) #column row counts
```

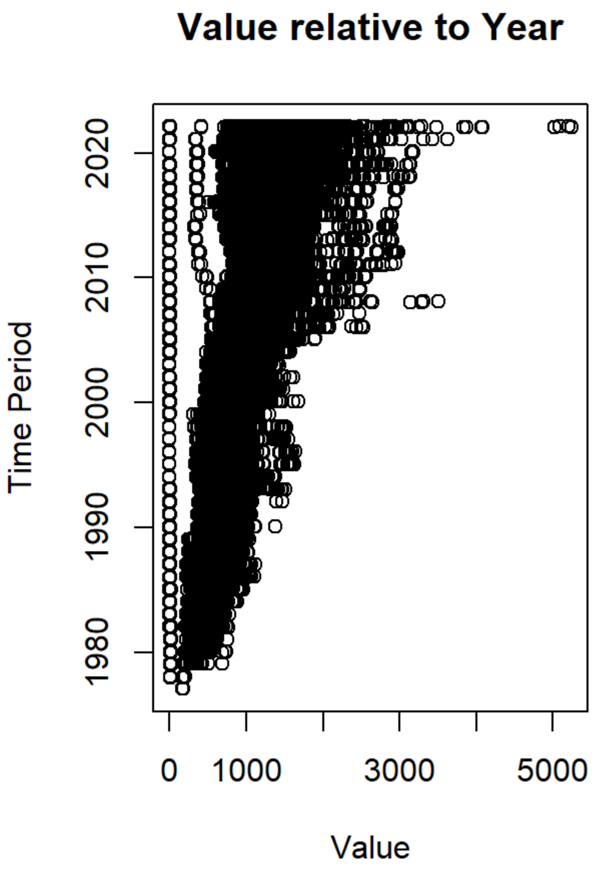
```
## 'data.frame':    74306 obs. of  6 variables:
##  $ Series_reference: chr  "CEPQ.S2371" "CEPQ.S2371" "CEPQ.S2371" "CEPQ.S2371" ...
##  $ Period          : num  1996 1997 1997 1997 1997 ...
##  $ Data_value      : num  899 884 925 932 929 940 956 958 970 1000 ...
##  $ Subject         : Factor w/ 4 levels "Capital Goods Price Index - CEP",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Group           : Factor w/ 18 levels "Energy Price Indexes - Base Period December 1996 quarter (=100
0)",..: 15 15 15 15 15 15 15 15 15 15 ...
##  $ Series_title_1  : Factor w/ 471 levels "Accident and health insurance services",..: 186 186 186 186 1
86 186 186 186 186 186 ...
```

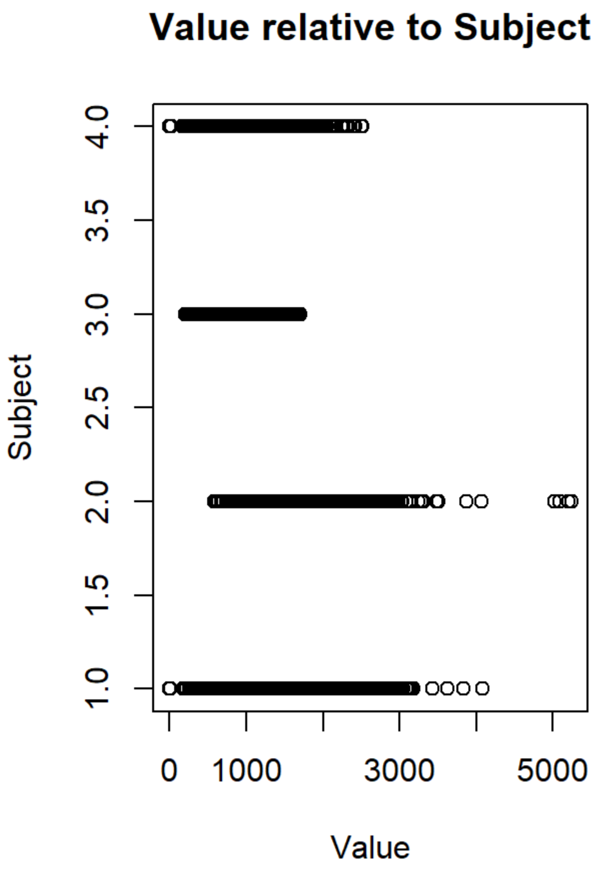
Informative graphs

```
par(mfrow=c(1,2))
cdplot(train$Subject~train$Period, main = "CD Plot - Subject~Period", ylab = "Subject", xlab = "Period")

hist(train$Period, main = "Business Price Index Year")
```



```
plot(train$Data_value, train$Period, main = "Value relative to Year", ylab = "Time Period", xlab = "Value")

plot(train$Data_value, train$Subject, main = "Value relative to Subject", ylab = "Subject", xlab = "Value")
```





Logistical Regression model

```
glm1 <- glm(Subject~Data_value+Period, data=train, family="binomial")
summary(glm1)
```

```
##
## Call:
## glm(formula = Subject ~ Data_value + Period, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1293   0.2821   0.3570   0.4286   5.3540
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.658e+02  3.987e+00  -66.66   <2e-16 ***
## Data_value  -4.364e-03  6.328e-05  -68.96   <2e-16 ***
## Period       1.357e-01  2.012e-03   67.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41531  on 59443  degrees of freedom
## Residual deviance: 34467  on 59441  degrees of freedom
## AIC: 34473
##
## Number of Fisher Scoring iterations: 5
```

Explanation of Logistic Regression Model Summary: The Logistic Regression Model summary shows the ‘deviance residuals’ which quantifies a given point’s contribution to the overall likelihood. Our deviance residuals -4 for Min, 0.28 for IQ, 0.357 for Median, 0.429 for 3, and 5.35 for Max.

Our coefficients shows the difference in the log odds of the target variable, which in this case the target variable is Subject with the Predictors being Data_value and Period.

Our null deviance shows us the lack of fit of the model considering only the intercept which in this case is 41531 on 59443 degrees of freedom. And our residual deviance shows us the lack of fit for our entire model which is 34467 on 59441 degrees of freedom. Our residual deviance is lower than our Null deviance which is good and helps show if our model is good.

Our AIC score is 34473 which is very bad, as we always want to see our AIC sore low.

Naive Bayes Model

```
library(e1071)
nb1 <- naiveBayes(Subject~Data_value+Period, data=train)
nb1
```

```
## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
## 
## A-priori probabilities:
## Y
## Capital Goods Price Index - CEP         Energy Statistics - NRG 
##                      0.11134850                      0.01766368 
##               Farm Inputs - FPI     Producers Price Index - PPI 
##                      0.24133638                      0.62965144 
## 
## Conditional probabilities:
##                                  Data_value
## Y                                      [,1]     [,2]
##   Capital Goods Price Index - CEP 1129.8948 429.2860
##   Energy Statistics - NRG         1636.4514 677.2911
##   Farm Inputs - FPI                835.2564 264.0013
##   Producers Price Index - PPI      966.4191 246.4744
## 
##                                  Period
## Y                                     [,1]      [,2]
##   Capital Goods Price Index - CEP 2005.051 10.960917
##   Energy Statistics - NRG         2005.858  9.603664
##   Farm Inputs - FPI               2005.475 11.532879
##   Producers Price Index - PPI     2009.813  8.403111
```

Naive Bayes Model Explanation:

This model shows that the Prior for the different factors of Subject, called A-priori. The priors are 0.111 for Capital Good Price Index, 0.017 for Energy Statistics, 0.24 for Farm Inputs, and 0.629 for Producers Price Index.

The likelihood data is shown under conditional probabilities and since our two predictors, Data_value and Period, are continuous we are shown the mean for each factor. The means for ‘Period’ are very similar, which means that looking at ‘Period’ alone doesn’t tell us much, but the Data_value values differ widely meaning that looking at Data_value can tell us something about the data..

Evaluating on Test Data with Classification Metrics - Logistical Regression

Accuracy

```
probs <- predict(glm1, newdata=test, type="response")
pred <- ifelse(probs>0.5, 2, 1)
acc1 <- mean(pred==as.integer(test$Subject))
print(paste("glm1 accuracy = ", acc1))
```

```
## [1] "glm1 accuracy =  0.0191091373973893"
```

```
table(pred, as.integer(test$Subject))
```

```
##
## pred    1    2    3    4
##    1  149  144   47   17
##    2 1487  135 3559 9324
```

Confusion Matrix

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
#adding this is the only was I can get a Confusion matrix
#Im making it a factor for the Matrix
test$Subject <- as.integer(test$Subject)
confusionMatrix(as.factor(pred), as.factor(test$Subject))
```

```
## Warning in confusionMatrix.default(as.factor(pred), as.factor(test$Subject)):
## Levels are not in the same order for reference and data. Refactoring data to
## match.
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2    3    4
##          1  149  144   47   17
##          2 1487  135 3559 9324
##          3    0    0    0    0
##          4    0    0    0    0
##
## Overall Statistics
##
##                Accuracy : 0.0191
##                  95% CI : (0.017, 0.0214)
##     No Information Rate : 0.6285
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0019
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity           0.09108 0.483871   0.0000   0.0000
## Specificity           0.98427 0.014606   1.0000   1.0000
## Pos Pred Value        0.41737 0.009307      NaN      NaN
## Neg Pred Value        0.89748 0.596639   0.7574   0.3715
## Prevalence            0.11008 0.018773   0.2426   0.6285
## Detection Rate        0.01003 0.009084   0.0000   0.0000
## Detection Prevalence  0.02402 0.975979   0.0000   0.0000
## Balanced Accuracy     0.53767 0.249239   0.5000   0.5000
```

ROC Curve and AUC Curve

```
#library(ROCR)
#p <- predict(glm1, newdata=test, type="response")
#pr <- prediction(p, test$Subject)
#prf <- performance(pr, measure = "tpr", x.measure = "fpr")
#plot(prf)

#auc <- performance(pr, measure = "auc")
#auc <- auc@y.values[[1]]
#auc
```

This is how I would get the ROC and AUC Curve. but my factors are not 2 levels. My smallest one is 4 levels, so I can’t get the plots. I get this error: Error in prediction(p, test$Subject) : Number of classes is not equal to 2. ROCR currently supports only evaluation of binary classification tasks.

Evaluating on Test Data with Classification Metrics - Naive Bayes

```
p1 <- predict(nb1, newdata=test, type="class")
table(p1, test$Subject)
```

```
##
## p1                                1    2    3    4
##   Capital Goods Price Index - CEP  207   39   16  218
##   Energy Statistics - NRG           29   81    0    1
##   Farm Inputs - FPI                239   34 1085  819
##   Producers Price Index - PPI     1161  125 2505 8303
```

```
mean(p1==test$Subject)
```

```
## [1] 0
```

Extracting Raw Probabilities

```
p1_raw <- predict(nb1, newdata=test, type="raw")
head(p1_raw)
```

```
##      Capital Goods Price Index - CEP Energy Statistics - NRG Farm Inputs - FPI
## [1,]                      0.09616368             0.006131783         0.3347769
## [2,]                      0.09301452             0.006165431         0.3036164
## [3,]                      0.09291227             0.006487293         0.2667319
## [4,]                      0.08261063             0.005772317         0.2630002
## [5,]                      0.07539140             0.005293504         0.2582000
## [6,]                      0.07724407             0.005500097         0.2478400
##      Producers Price Index - PPI
## [1,]                   0.5629277
## [2,]                   0.5972037
## [3,]                   0.6338686
## [4,]                   0.6486169
## [5,]                   0.6611151
## [6,]                   0.6694158
```

Strength and Weaknesses of Logistic Regression and Naive Bayes

The strengths of Logistic Regression is that it has a good probability output, is computationally inexpensive, and If the classes are linearly separable it separates them well. The weakness of linear regression is that it has a tendency to under fit the data and is not flexible enough to display non-linear decision boundaries. The strengths of Naïve Bayes models are that they work very well with small data sets, they're easy to both interpret and implement, and they handle high dimensions well. The weaknesses of Naïve Bayes are that the model will assume that predictors are all independent of one another so if they are not independent of one another, the model's performance will falter. Other weaknesses is that the model is not suited for larger data sets, and the model makes guesses for values in the test set that didn't occur in the training data.

Classification metrics benefits, drawbacks, and information

Logistic Regression Metrics:

Accuracy: The Accuracy metric shows the percentage of correct predictions relative to the total number of examples. The benefit of this metric is that is show us how accurate our data predictions are. However, this metric is determined by the number of correct predictions total and does not separate based different classes predicted, but rather gives the total picture. In this case, out accuracy score was 0.019, meaning out data predictions were not accurate.

Confusion Matrix: The Confusion Matrix displays a table of predictors and their True Values. The values shown are True Positive, False Positive, False Negative, and True Negative. The benefit of this metric is that it allows you to see how many correct and wrong predictions you data made. The drawback is that if the data being modeled is unbalanced, the accuracy could to be high even if the model is inaccurate due to the fact that the model could predict all the data points towards the majority class. This would make the accuracy high even if the model is bad. The confusion Matrix shows that our accuracy is very low and that our model had a low amount of True Positives.

Sensitivity and Specificity: The Sensitivity measures the true positive rate of the data, meaning how many times did the model correctly guess positive (TP) over the amount of data points that were positive (TP + FN). The Specificity on the other hand measures the true negative rate of the data, meaning how many times did the model correctly guess negative (TN) over the amount of data points that were actually negative (TN + FP). The benefit of these metrics is that they allow a closer look into how good the data model is, but the drawback is once again that their values could be high even if the model is bad if the data is unbalanced. This drawback can be seen in the logistic regression's metrics, as the specificity and sensitivity metrics don't accurately depict how bad the model is.

Kappa: The Kappa adjusts accuracy to account for correct prediction by chance, and the Kappa depicted is -0.0019 meaning that our Kappa score is a poor agreement. -0.0019

Naive Bayes Metrics *Class Probabilities*: This metric shows the class predictions, and the table shows us the class probabilities for all the four factors of Subject. Our table shows us that the class predictions skew in favor of the FPI and PPI class, with the first two classes having less predictions.

Raw Probabilities: This metric shows the raw probabilities from the predictions and shows us the percentage for each of the four factors of Subject. The raw probabilities for 'EP and PPI are the highest, with CEP's being close to 1 and FPI and NRG's probabilities are the lowest with NRG's being almost at 0.