Logistic Regression Analysis of Banknotes Authenticity

C M Nafi and Sneha Verma

1/28/2021

## Abstract

This research focuses on whether a banknote is authentic or not based on images that were taken and its features. We started by conducting basic exploration of each of the variables to understand if there were any skewness that needed to be dealt with. There was low to moderate skewness but it was not of a large concern due to which we did not conduct any transformations on our variables. Then, we conducted a first-order model to see which variables are signficant and found that all variables except entropy are signficant in determining if the image is of a real or fake banknote. After this, we conducted step-wise regression and got similar results because of which we added interaction effects and did stepwise regression with the AIC criterion and found that the interaction between variance and entropy, and kurtosis and entropy are significant at a 0.05 level. It appears that our model is a close-to-perfect fit of the dataset because it has linearity, as checked from model diagnostics, and high sensitivity and specificity rates indicating that its predictive powers is high. Further, the area under the ROC curve is close to 1 (0.9998), further confirmation of the model’s perfect fitness. There are points that have high leverage values and there is one point that has a high Cook’s distance value. For further investigation, we would suggest looking at these points to understand how we can deal with them.

NOTE: Before running the file, make sure that the location of the data corresponds with the location on your device.

This dataset (obtained from [Kaggle](https://www.kaggle.com/ritesaluja/bank-note-authentication-uci-data)) was generated from genuine and forged banknote images.An industrial camera was used for print inspection and the Wavelet Transform tool were used to extract features from the images.

This research project will focus on the central question:

**How well can the image features predict if the banknote is counterfeit?**

The dataset has 5 features and 1372 observations. The features are as follows:

1. **variance**: is a continuous value which finds how each pixel varies from the neighboring pixels and classifies them into different regions. If the pixel values are close to the mean pixel value then the variance is low and if they are not close to the mean pixel value theen it is a high variance.
2. **skewness**: is a continuous value measuring the lack of symmetry. A symmetric data would have a value of 0. Negative values indicate left-skewness and positive values indicate right-skewness.
3. **kurtosis**: a continuous value measuring whether the data is heavy-tailed or light-tailed relative to a normal distribution. Datasets with high kurtosis have heavy-tails or outliers and those with low kurtosis have light tails or lack of outliers.
4. **entropy**: is a continuous value representing a quantity used to describe the amount of information which must be coded for, by a compression algorithm. A higher entropy value implies that there is a higher amount of information in the image.
5. **class**: is the target, binary variable where a value of 0 represents a geunine or authentic bank note and a value of 1 represents a fake note.

## Import Dataset

banknote = read.csv('Original\_Banknote\_Authentication.csv')  
head(banknote)

## variance skewness curtosis entropy class  
## 1 3.62160 8.6661 -2.8073 -0.44699 0  
## 2 4.54590 8.1674 -2.4586 -1.46210 0  
## 3 3.86600 -2.6383 1.9242 0.10645 0  
## 4 3.45660 9.5228 -4.0112 -3.59440 0  
## 5 0.32924 -4.4552 4.5718 -0.98880 0  
## 6 4.36840 9.6718 -3.9606 -3.16250 0

# Let us rename the variable curtosis to the more commonly used spelling  
names(banknote)[names(banknote) == 'curtosis'] = 'kurtosis'  
names(banknote)[names(banknote) == 'class'] = 'notAuthentic'

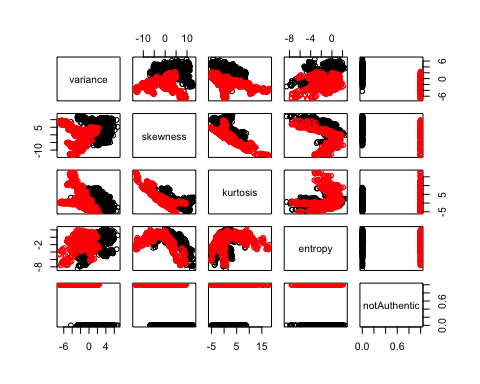
## Data Characteristics

Looking at this dataset, it can be observed that there are five variables (one of them is the target variable: notAuthentic).

### Scatterplot Matrix

Let us create a scatterplot matrix coloured by the outcome variable:

pairs(banknote, col=banknote$notAuthentic+1)



The scatterplot matrix shows that there might be a relationship between some of the predictor variables and the response variable. It appears that for lower values of variance in the image, the banknote is likely to be fake and images with higher variance values were of genuine bank notes. Further, it appears that genuine bank notes have higher skewness; however, there does not appear to be a high level of correlation because fake bank notes also have relatively high bank skewness. Similarly, genuine bank notes have lower values of kurtosis as compared to fake notes; however, fake notes have kurtosis values of the entire range, this may make it harder to differentiate between genuine and fake banknotes. Lastly, it appears that entropy is not a good predictor of authenticity because images of fake and genuine banknotes have entropy values of the entire range of possible values indicating a weak relationship between teh two variables.

### Correlation Matrix

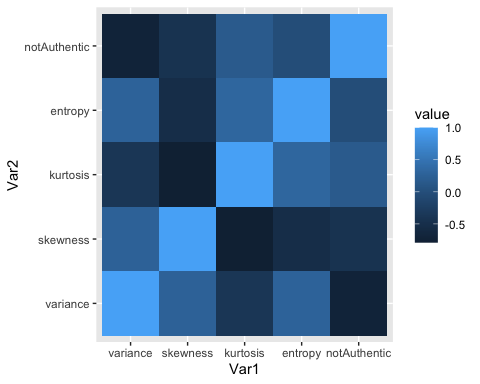
#install.packages('reshape2')  
#install.packages('ggplot2')  
  
banknoteCorr <- round(cor(banknote),2)  
head(banknoteCorr)

## variance skewness kurtosis entropy notAuthentic  
## variance 1.00 0.26 -0.38 0.28 -0.72  
## skewness 0.26 1.00 -0.79 -0.53 -0.44  
## kurtosis -0.38 -0.79 1.00 0.32 0.16  
## entropy 0.28 -0.53 0.32 1.00 -0.02  
## notAuthentic -0.72 -0.44 0.16 -0.02 1.00

library(reshape2)  
melted\_banknote <- melt(banknoteCorr)  
head(melted\_banknote)

## Var1 Var2 value  
## 1 variance variance 1.00  
## 2 skewness variance 0.26  
## 3 kurtosis variance -0.38  
## 4 entropy variance 0.28  
## 5 notAuthentic variance -0.72  
## 6 variance skewness 0.26

library(ggplot2)  
ggplot(data = melted\_banknote, aes(x=Var1, y=Var2, fill=value)) +   
 geom\_tile()



This matrix shows the correlation matrix between the variables in the data set. Lighter boxes indicate higher correlation values.

It appears that there is a high correlation between the authenticity of banknotes and the kurtosis. There also appears to be a relatively higher correlation between authenticity and entropy compared to the correlation between skewness and variance with authenticity. This is interesting because the scatterplot matrix between authenticity and entropy does not show a strong relationship.

### Plots of Individual Variables

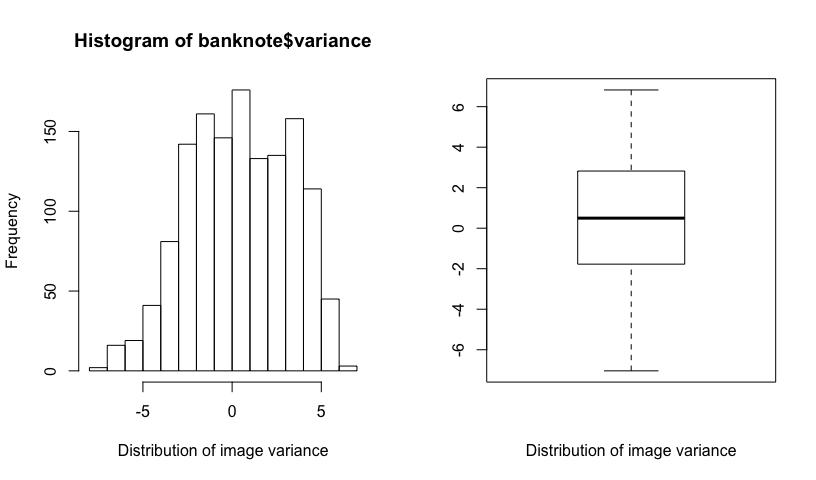
#### variance

summary(banknote$variance)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -7.0421 -1.7730 0.4962 0.4337 2.8215 6.8248

The minimum value of variance is -7 variance units and the maximum value is 6.8 variance units with a median of 0.496 and a mean of 0.434.

par(mfrow=c(1,2))  
hist(banknote$variance, xlab = "Distribution of image variance")  
boxplot(banknote$variance, xlab = "Distribution of image variance")



The histogram of the variance shows mild left-skewness; however, since it the skewness is only slight, no transformation will be conducted. Further, the boxplot shows that there are no outliers in this variable.

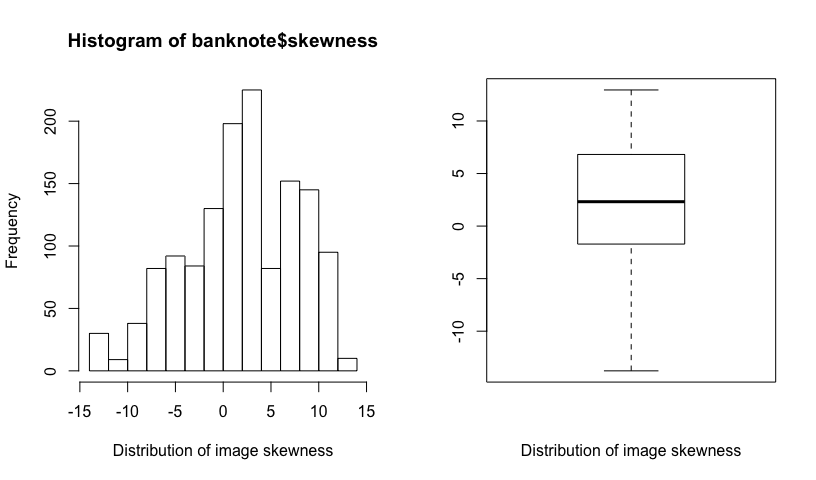
#### skewness

summary(banknote$skewness)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -13.773 -1.708 2.320 1.922 6.815 12.952

The minimum value of skewness is -13.77 skewness units and the maximum value is 12.95 skewness units with a median of 2.32 and a mean of 1.92.

par(mfrow=c(1,2))  
hist(banknote$skewness, xlab = "Distribution of image skewness")  
boxplot(banknote$skewness, xlab = "Distribution of image skewness")



The histogram of the skewness shows slight or moderate left-skewness; however, since it the skewness is only slight, no transformation will be conducted. Further, the boxplot shows that there are no outliers in this variable. The boxplot also shows the left-skewness with a long lower tail.

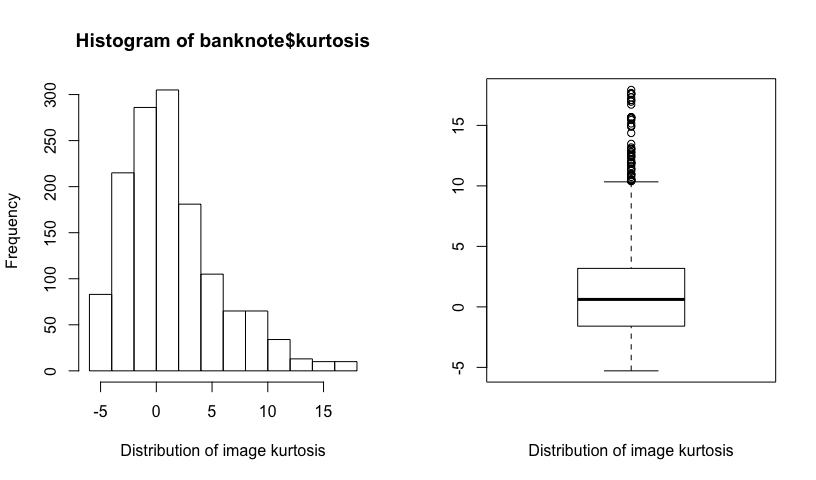
#### kurtosis

summary(banknote$kurtosis)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -5.2861 -1.5750 0.6166 1.3976 3.1793 17.9274

The minimum value of variance is -5.29 kurtosis units and the maximum value is 17.93 kurtosis units with a median of 0.617 and a mean of 1.398.

par(mfrow=c(1,2))  
hist(banknote$kurtosis, xlab = "Distribution of image kurtosis")  
boxplot(banknote$kurtosis, xlab = "Distribution of image kurtosis")



The histogram of the variance shows moderate right-skewness. Further, the boxplot displays this skewness with a longer upper tail and also displays outliers at the upper end of the distribution. However, since there is moderate skewness, we will not transform the variable at this point.

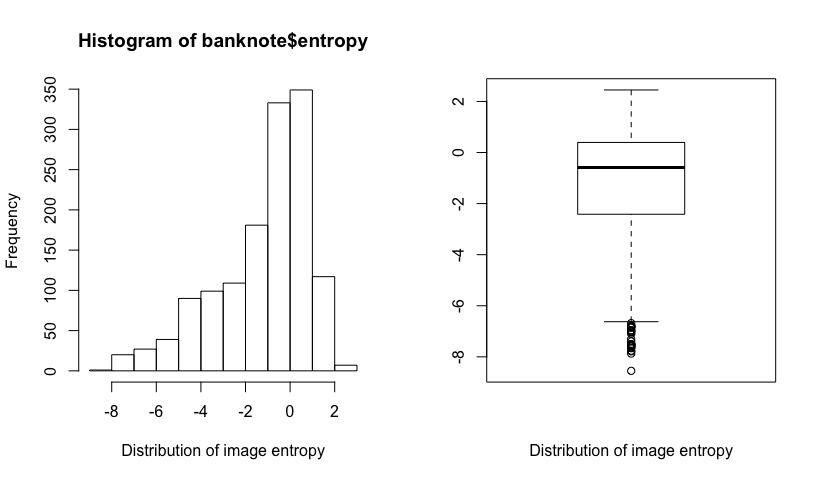
#### entropy

summary(banknote$entropy)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -8.5482 -2.4135 -0.5867 -1.1917 0.3948 2.4495

The minimum value of variance is -8.55 entropy units and the maximum value is 2.45 entropy units with a median of -0.59 and a mean of -1.19.

par(mfrow=c(1,2))  
hist(banknote$entropy, xlab = "Distribution of image entropy")  
boxplot(banknote$entropy, xlab = "Distribution of image entropy")



The histogram of the variance shows moderate left-skewness. Further, the boxplot displays this skewness with a long lower tail and also displays outliers at the lower end of the distribution. However, since there is only moderate skewness, we will not apply a transformation at the point.

## First-Order Logistic Regression Model

Let us now fit a logistic regression model that includes all of the variables.

fit1 = glm(notAuthentic ~ variance + skewness + kurtosis + entropy, data = banknote, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(fit1)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis +   
## entropy, family = binomial, data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.70001 0.00000 0.00000 0.00029 2.24614   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.3218 1.5589 4.697 2.64e-06 \*\*\*  
## variance -7.8593 1.7383 -4.521 6.15e-06 \*\*\*  
## skewness -4.1910 0.9041 -4.635 3.56e-06 \*\*\*  
## kurtosis -5.2874 1.1612 -4.553 5.28e-06 \*\*\*  
## entropy -0.6053 0.3307 -1.830 0.0672 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 49.891 on 1367 degrees of freedom  
## AIC: 59.891  
##   
## Number of Fisher Scoring iterations: 12

While this fit produces a warning message, it does converge on parameter estimates and p-values. From the resultant summary, we can see that all variables are significant at the 0.001 significance level, except for entropy which is significant at the 0.1 level only. The slopes are given in the logit form, let us convert it to the odds ratio.

exp(fit1$coefficients)

## (Intercept) variance skewness kurtosis entropy   
## 1.512932e+03 3.861323e-04 1.513170e-02 5.054731e-03 5.459003e-01

exp(confint(fit1))

## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## 2.5 % 97.5 %  
## (Intercept) 1.298651e+02 6.854946e+04  
## variance 5.862477e-06 6.212265e-03  
## skewness 1.708813e-03 6.405556e-02  
## kurtosis 3.045858e-04 3.203827e-02  
## entropy 2.730819e-01 1.038367e+00

The odds of the note being fake is 0.00039 times higher for each unit of variance with a 95% confidence interval of 0.0000059 to 0.00621 higher odds of the note being fake per variance unit.

The odds of the note being fake is 0.0151 times higher for each unit of skewness with a 95% confidence interval of 0.00171 to 0.0641 higher odds of the note being fake per skewness unit.

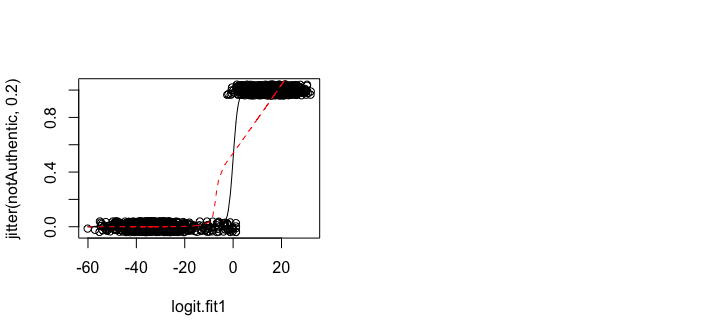
The odds of the note being fake is 0.00505 times higher for each unit of kurtosis with a 95% confidence interval of 0.000306 to 0.0320 higher odds of the note being fake per kurtosis unit.

The odds of the note being fake is 0.546 times higher for each unit of entropy with a 95% confidence interval of 0.273 to 1.04 higher odds of the note being fake per entropy unit.

### Model Diagnostics

*1) Jittered response v/s predicted values*

par(mfrow=c(1,2))  
logit.fit1 = predict(fit1)  
plot (jitter (notAuthentic, 0.2) ~ logit.fit1, data=banknote)  
  
pihat.fit1 = predict(fit1, type='response')  
pihat.ord = order(pihat.fit1)  
lines (logit.fit1[pihat.ord], pihat.fit1[pihat.ord])  
lines (lowess (logit.fit1[pihat.ord], pihat.fit1[pihat.ord]), col='red', lty=2)



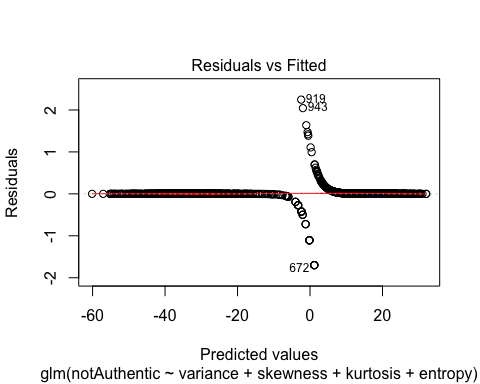
From the “Jittered response v/s predicted values” plot, we can see the probability of predicted logit. We can see that the probability is 1 when logit values are greater than 0. This plot shows that the logistic regression is a perfect fit to the model because the logit values when the note is genuine are clustered at the lower-left of the graph and the values when the note is fake are clustered at the top-right and the logistic curve predicts this clear distinction and nealy perfectly fits the data.

On the other hand, the smoothing spline bends at certain data points which is not mirrored by the logistic curve.

We are going to check the residual v/s Fitted values next to check the linearity.

*2) Residuals v/s Fitted Values*

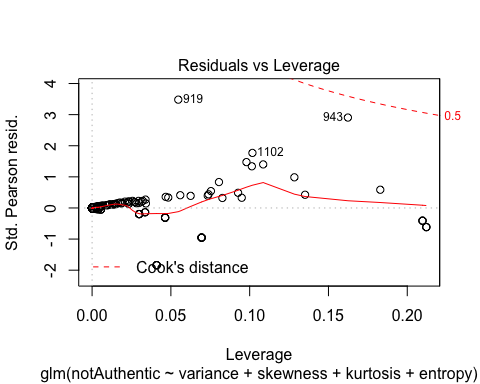
plot(fit1, which = 1)



The trendline in the residual v/s fitted plot shows little to no deviation from the horizontal zero line showing that the distribution of the residuals is not concerning; it shows clear linearity. Further, there do appear to be a few points with relatively high outliers (line observations 919 and 943).

*3) Leverage*

plot(fit1, which = 5)



In this ‘Residuals vs Leverage plot’, we are trying to check unusual points in this dataset. On the residuals vs leverage plot, there is a residual greater than +/- 3 standardized pearson residuals and one residual that is close to a value of 3.

Let’s calculate very high values for Leverage in our dataset: Using the formula (3(k+1))/n we found that our very high leverage value is 0.01093. There are many data points which have high leverage points in this plot. Thus, there are points with high amounts of influence.

We can also see the Cook’s distance on the plot, which looks fine because all of the values are less than 0.5.

*4) Multicollinearity*

car::vif(fit1)

## variance skewness kurtosis entropy   
## 74.744334 215.208045 407.716399 2.281083

We can see that some predictors are highly correlated to each other because their VIF value are greater than 5. Only entropy has a VIF value of less than 5.

## Model Selection

### Stepwise Regression

Let us now apply stepwise regression to the first-order model to see which variables should be kept in the model.

fit1aic = step(fit1, direction = 'both')

## Start: AIC=59.89  
## notAuthentic ~ variance + skewness + kurtosis + entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 49.89 59.89  
## - entropy 1 53.30 61.30  
## - skewness 1 636.52 644.52  
## - kurtosis 1 719.24 727.24  
## - variance 1 1145.48 1153.48

summary(fit1aic)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis +   
## entropy, family = binomial, data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.70001 0.00000 0.00000 0.00029 2.24614   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.3218 1.5589 4.697 2.64e-06 \*\*\*  
## variance -7.8593 1.7383 -4.521 6.15e-06 \*\*\*  
## skewness -4.1910 0.9041 -4.635 3.56e-06 \*\*\*  
## kurtosis -5.2874 1.1612 -4.553 5.28e-06 \*\*\*  
## entropy -0.6053 0.3307 -1.830 0.0672 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 49.891 on 1367 degrees of freedom  
## AIC: 59.891  
##   
## Number of Fisher Scoring iterations: 12

According to this result, there are no steps that can be taken to decrease the value of AIC.

Let us now try stepwise regression with the BIC criterion:

fit1bic = step(fit1, direction = "both", k = log(fit1$rank + fit1$df.residual))

## Start: AIC=86.01  
## notAuthentic ~ variance + skewness + kurtosis + entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - entropy 1 53.30 82.19  
## <none> 49.89 86.01  
## - skewness 1 636.52 665.42  
## - kurtosis 1 719.24 748.14  
## - variance 1 1145.48 1174.38

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=82.19  
## notAuthentic ~ variance + skewness + kurtosis

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 53.30 82.19  
## + entropy 1 49.89 86.01  
## - kurtosis 1 722.03 743.70  
## - skewness 1 850.17 871.84  
## - variance 1 1399.79 1421.46

summary(fit1bic)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis,   
## family = binomial, data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.62625 0.00000 0.00000 0.00092 2.48704   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.8850 1.3838 4.975 6.52e-07 \*\*\*  
## variance -6.7835 1.3950 -4.863 1.16e-06 \*\*\*  
## skewness -3.5067 0.6932 -5.059 4.22e-07 \*\*\*  
## kurtosis -4.4642 0.9006 -4.957 7.16e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 53.299 on 1368 degrees of freedom  
## AIC: 61.299  
##   
## Number of Fisher Scoring iterations: 12

The result in BIC stepwise regression removed entropy from the model.  
We can also see that the residual deviance is slightly better in AIC (49.891) compare to BIC (53.299) although the diffeerence is not too large. Due to this, if we were to continue with the model, we would recommend using the AIC criterion.

### Interaction Model

Now we are going to check interaction effects for all variables:

fit\_int1 = glm(notAuthentic ~ (variance + skewness + kurtosis + entropy)^2, family = binomial, data = banknote)

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(fit\_int1)

##   
## Call:  
## glm(formula = notAuthentic ~ (variance + skewness + kurtosis +   
## entropy)^2, family = binomial, data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -9.878e-05 -2.100e-08 -2.100e-08 2.100e-08 1.633e-04   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 29.5434 8047.5468 0.004 0.997  
## variance -27.5974 11001.8523 -0.003 0.998  
## skewness -14.9888 9645.5015 -0.002 0.999  
## kurtosis -24.1502 6022.3385 -0.004 0.997  
## entropy 15.1990 12929.0499 0.001 0.999  
## variance:skewness 1.9539 4227.3918 0.000 1.000  
## variance:kurtosis -1.7149 2004.8976 -0.001 0.999  
## variance:entropy 6.2379 2116.2625 0.003 0.998  
## skewness:kurtosis -0.1852 1466.5631 0.000 1.000  
## skewness:entropy 0.4928 2245.0735 0.000 1.000  
## kurtosis:entropy 3.7040 3172.2060 0.001 0.999  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1.8851e+03 on 1371 degrees of freedom  
## Residual deviance: 1.1107e-07 on 1361 degrees of freedom  
## AIC: 22  
##   
## Number of Fisher Scoring iterations: 25

This interaction model gives an error that the algorithm has not been able to converge and all the variables in this model have high p-values close to 1.0 making them insignificant.

Let us try another interaction plot with a few of the variables:

fit\_int2 = glm (notAuthentic ~ variance + skewness + kurtosis + entropy +   
 entropy:(variance + skewness + kurtosis), data=banknote, family=binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(fit\_int2)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis +   
## entropy + entropy:(variance + skewness + kurtosis), family = binomial,   
## data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.66356 0.00000 0.00000 0.00011 1.79125   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.5079 2.3175 4.103 4.08e-05 \*\*\*  
## variance -10.7681 2.6582 -4.051 5.10e-05 \*\*\*  
## skewness -5.2287 1.3067 -4.001 6.30e-05 \*\*\*  
## kurtosis -6.9477 1.7420 -3.988 6.65e-05 \*\*\*  
## entropy -0.3143 0.5467 -0.575 0.5653   
## variance:entropy -1.0563 0.4487 -2.354 0.0186 \*   
## skewness:entropy -0.2954 0.2522 -1.171 0.2415   
## kurtosis:entropy -0.5530 0.3362 -1.645 0.1000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 42.872 on 1364 degrees of freedom  
## AIC: 58.872  
##   
## Number of Fisher Scoring iterations: 13

This model converges to a solution. With the results table, one can observe that all the variables (without interactions) are significant, except entropy and the variance and entropy interaction are significant at the 0.05 significance level.

Before we interpret the results, let us attempt this model with centered variables:

banknote$variance.c = banknote$variance - mean(banknote$variance)  
  
banknote$skewness.c = banknote$skewness - mean(banknote$skewness)  
  
banknote$kurtosis.c = banknote$kurtosis - mean(banknote$kurtosis)  
  
banknote$entropy.c = banknote$variance - mean(banknote$entropy)

fit\_int\_c\_1 = glm (notAuthentic ~ variance.c + skewness.c + kurtosis.c + entropy.c +   
 entropy.c:(variance.c + skewness.c + kurtosis.c), data=banknote, family=binomial)

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(fit\_int\_c\_1)

##   
## Call:  
## glm(formula = notAuthentic ~ variance.c + skewness.c + kurtosis.c +   
## entropy.c + entropy.c:(variance.c + skewness.c + kurtosis.c),   
## family = binomial, data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.141e-04 -2.000e-08 -2.000e-08 2.000e-08 3.313e-04   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -81.5015 9620.7003 -0.008 0.993  
## variance.c -52.3385 6806.6604 -0.008 0.994  
## skewness.c -33.5508 5534.6781 -0.006 0.995  
## kurtosis.c -43.0814 8092.3435 -0.005 0.996  
## entropy.c NA NA NA NA  
## variance.c:entropy.c 3.4520 2016.5412 0.002 0.999  
## skewness.c:entropy.c 4.0727 3188.0629 0.001 0.999  
## kurtosis.c:entropy.c 0.3897 5229.9786 0.000 1.000  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1.8851e+03 on 1371 degrees of freedom  
## Residual deviance: 2.3448e-07 on 1365 degrees of freedom  
## AIC: 14  
##   
## Number of Fisher Scoring iterations: 25

Looking at the results, we can see that this model failed to converge and none of the variables are signficant (possibly due to multicollinaerity).

Let us use the fit\_int2 model and conduct stepwise regression:

fit\_int2\_aic = step(fit\_int2, direction = 'both')

## Start: AIC=58.87  
## notAuthentic ~ variance + skewness + kurtosis + entropy + entropy:(variance +   
## skewness + kurtosis)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - skewness:entropy 1 43.657 57.657  
## - kurtosis:entropy 1 44.512 58.512  
## <none> 42.872 58.872  
## - variance:entropy 1 46.396 60.396

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=57.66  
## notAuthentic ~ variance + skewness + kurtosis + entropy + variance:entropy +   
## kurtosis:entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 43.66 57.66  
## + skewness:entropy 1 42.87 58.87  
## - kurtosis:entropy 1 49.02 61.02  
## - variance:entropy 1 49.29 61.29  
## - skewness 1 583.35 595.35

summary(fit\_int2\_aic)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis +   
## entropy + variance:entropy + kurtosis:entropy, family = binomial,   
## data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.67104 0.00000 0.00000 0.00012 1.90054   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.0736 2.1492 4.222 2.42e-05 \*\*\*  
## variance -10.2034 2.4411 -4.180 2.92e-05 \*\*\*  
## skewness -4.7746 1.1018 -4.334 1.47e-05 \*\*\*  
## kurtosis -6.4497 1.5309 -4.213 2.52e-05 \*\*\*  
## entropy -0.6800 0.4298 -1.582 0.11362   
## variance:entropy -0.6804 0.2341 -2.906 0.00366 \*\*   
## kurtosis:entropy -0.2189 0.1007 -2.174 0.02969 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 43.657 on 1365 degrees of freedom  
## AIC: 57.657  
##   
## Number of Fisher Scoring iterations: 13

The AIC criterion keeps all of the non-interacted variables and two of the interactions (variance and entropy, and kurtosis and entropy).

The residual deviance of this model is 43.657 units which is a bit lower than the residual deviance from the non-interacted model.

Let us have a look at the stepwise regression with the BIC criterion:

fit\_int2bic = step(fit\_int2, direction = "both", k = log(fit\_int2$rank + fit\_int2$df.residual))

## Start: AIC=100.66  
## notAuthentic ~ variance + skewness + kurtosis + entropy + entropy:(variance +   
## skewness + kurtosis)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - skewness:entropy 1 43.657 94.225  
## - kurtosis:entropy 1 44.512 95.080  
## - variance:entropy 1 46.396 96.964  
## <none> 42.872 100.664

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=94.22  
## notAuthentic ~ variance + skewness + kurtosis + entropy + variance:entropy +   
## kurtosis:entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - kurtosis:entropy 1 49.02 92.36  
## - variance:entropy 1 49.29 92.64  
## <none> 43.66 94.22  
## + skewness:entropy 1 42.87 100.66  
## - skewness 1 583.35 626.69

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=92.36  
## notAuthentic ~ variance + skewness + kurtosis + entropy + variance:entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - variance:entropy 1 49.89 86.01  
## <none> 49.02 92.36  
## + kurtosis:entropy 1 43.66 94.22  
## + skewness:entropy 1 44.51 95.08  
## - skewness 1 583.93 620.06  
## - kurtosis 1 674.13 710.25

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=86.01  
## notAuthentic ~ variance + skewness + kurtosis + entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - entropy 1 53.30 82.19  
## <none> 49.89 86.01  
## + skewness:entropy 1 48.09 91.44  
## + variance:entropy 1 49.02 92.36  
## + kurtosis:entropy 1 49.29 92.64  
## - skewness 1 636.52 665.42  
## - kurtosis 1 719.24 748.14  
## - variance 1 1145.48 1174.38

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=82.19  
## notAuthentic ~ variance + skewness + kurtosis

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 53.30 82.19  
## + entropy 1 49.89 86.01  
## - kurtosis 1 722.03 743.70  
## - skewness 1 850.17 871.84  
## - variance 1 1399.79 1421.46

summary(fit\_int2bic)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis,   
## family = binomial, data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.62625 0.00000 0.00000 0.00092 2.48704   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.8850 1.3838 4.975 6.52e-07 \*\*\*  
## variance -6.7835 1.3950 -4.863 1.16e-06 \*\*\*  
## skewness -3.5067 0.6932 -5.059 4.22e-07 \*\*\*  
## kurtosis -4.4642 0.9006 -4.957 7.16e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 53.299 on 1368 degrees of freedom  
## AIC: 61.299  
##   
## Number of Fisher Scoring iterations: 12

The BIC criterion removes all interaction effects and removes entropy from the model. The residual deviance of this model is the same as the non-interacted model of 53.299 (since the resultant model is the same as well).

Since the residual deviance is smaller for the AIC model, we will use that as our final model.

Let us now have a look at the signficant interactions remaining in the model:

*1) Variance:entropy*

library(ggplot2)  
library(dplyr)

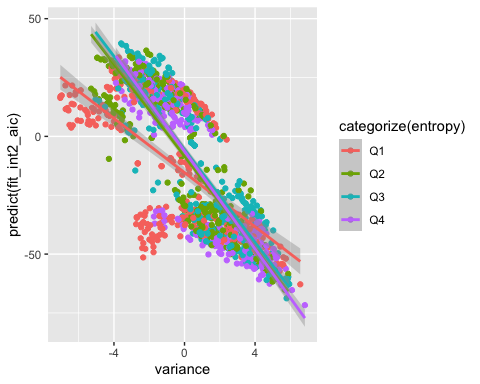
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

categorize = function (x) {  
 quartiles = summary (x) [c(2, 3, 5)]  
 result = rep ("Q1", length (x))  
 result [(quartiles[1] < x) & (x <= quartiles [2])] = "Q2"  
 result [(quartiles[2] < x) & (x <= quartiles [3])] = "Q3"  
 result [quartiles[3] < x] = "Q4"  
 return (result)  
}  
  
with(banknote,  
 qplot(x = variance, y = predict(fit\_int2\_aic), color = categorize(entropy)) +  
 geom\_smooth(method = 'lm'))

## `geom\_smooth()` using formula 'y ~ x'



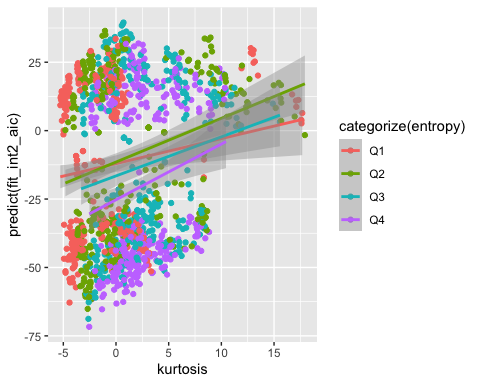
The slopes of variance and entropy are different for each category of entropy (based on quartiles). It shows that the logit prediction of the unauthenticity of banknotes decrease steeply with higher levels of variance as entropy increases. However, as the variance increases, the predicted logit fit of unauthenticity increases at a slower rate for the first quartile values of entropy.

Therefore, the relationship between variance and the unauthetnicity of banknotes is significant for different values of entropy.

1. *Kurtosis:entropy*

with(banknote,  
 qplot(x = kurtosis, y = predict(fit\_int2\_aic), color = categorize(entropy)) +  
 geom\_smooth(method = 'lm'))

## `geom\_smooth()` using formula 'y ~ x'



The slopes of kurtosis and entropy are different for each category of entropy (based on quartiles). The slopes for each category of entropy is not very different which is why the interaction is not extremely significant. The value of predicted logit values is the lowest for the largest quartile group of entropy as kurtosis increases. However, the difference in the slopes for the rest of the categories of entropy are not extremely different.

Therefore, the relationship between kurtosis and the unauthetnicity of banknotes is slightly significant for different values of entropy.

## Final Model

### Parameter Estimates

The final model is the AIC model of the interaction model (fit\_int2\_aic).

Let us have a look at the summary of the model one more time to understand the parameters:

summary(fit\_int2\_aic)

##   
## Call:  
## glm(formula = notAuthentic ~ variance + skewness + kurtosis +   
## entropy + variance:entropy + kurtosis:entropy, family = binomial,   
## data = banknote)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.67104 0.00000 0.00000 0.00012 1.90054   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.0736 2.1492 4.222 2.42e-05 \*\*\*  
## variance -10.2034 2.4411 -4.180 2.92e-05 \*\*\*  
## skewness -4.7746 1.1018 -4.334 1.47e-05 \*\*\*  
## kurtosis -6.4497 1.5309 -4.213 2.52e-05 \*\*\*  
## entropy -0.6800 0.4298 -1.582 0.11362   
## variance:entropy -0.6804 0.2341 -2.906 0.00366 \*\*   
## kurtosis:entropy -0.2189 0.1007 -2.174 0.02969 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1885.122 on 1371 degrees of freedom  
## Residual deviance: 43.657 on 1365 degrees of freedom  
## AIC: 57.657  
##   
## Number of Fisher Scoring iterations: 13

Looking at this model, we can see that all individual predictors, except entropy, are significant at the 0.001 level and the two interaction effects are also significant. However, the interaction between variance and entropy is more significant at the 0.01 level, while the interaction between kurtosis and entropy is signficant at the 0.05 level.

Since variance, kurtosis, and entropy are involved in interactions, it is harder to interpret the individual parameter estimates. However, we can interpret the estimate of skewness.

Looking at the results here, we can see that variance, skewness, kurtosis, and entropy have a negative effect on the banknote being not authentic. Further, both the interactions also have a negative effect on the response variable. We can also see that the residual deviance is 43.567 units.

### Odds Ratio

cbind.data.frame(exp.beta = exp(fit\_int2\_aic$coefficients), exp(as.data.frame(confint(fit\_int2\_aic))))

## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## exp.beta 2.5 % 97.5 %  
## (Intercept) 8.722345e+03 3.051646e+02 1.697963e+06  
## variance 3.704248e-05 1.006429e-07 1.777217e-03  
## skewness 8.441797e-03 5.861320e-04 4.827116e-02  
## kurtosis 1.581010e-03 3.867806e-05 1.771885e-02  
## entropy 5.066349e-01 1.967095e-01 1.123087e+00  
## variance:entropy 5.063919e-01 3.001213e-01 8.863459e-01  
## kurtosis:entropy 8.033825e-01 6.347978e-01 9.701447e-01

Looking at the table produced, we can interpret the odds ratio and confidence intervals for skewness which is the only variable not involved in an interaction (the interpretations of the interaction effects are shown above as interaction plots):

The odds of the banknote being fake is 0.00844 times higher for a one-unit increase in skewness. We are 95% confident that the odds of the banknote being fake is between 0.000586 and 0.0483 times larger for a one unit increase in skewness.

### Confidence and Prediction Intervals

fit.logits = predict (fit\_int2\_aic, se.fit=T)  
  
# Logit scale predictions and confidence limits (lwr, upr)  
  
fit.preds = data.frame (logit.fit = fit.logits$fit,  
se.fit = fit.logits$se.fit)  
fit.preds$logit.lwr = fit.preds$logit.fit - 1.96 \* fit.preds$se.fit  
fit.preds$logit.upr = fit.preds$logit.fit + 1.96 \* fit.preds$se.fit  
  
# Probability scale predictions and confidence limits (lwr, upr)  
  
fit.preds$prob.fit = exp (fit.preds$logit.fit) / (1 + exp (fit.preds$logit.fit))  
fit.preds$prob.lwr = exp (fit.preds$logit.lwr) / (1 + exp (fit.preds$logit.lwr))  
fit.preds$prob.upr = exp (fit.preds$logit.upr) / (1 + exp (fit.preds$logit.upr))  
###   
  
sort\_aic = order(as.data.frame(fit.preds)$prob.fit)  
as.data.frame(fit.preds)[sort\_aic,][c(230, 691, 606, 503, 859, 1286), c(5,6,7)]

## prob.fit prob.lwr prob.upr  
## 244 2.184343e-20 1.905303e-29 2.504250e-11  
## 217 4.616228e-10 2.066490e-14 1.031185e-05  
## 115 2.367758e-14 1.694054e-20 3.309386e-08  
## 107 3.354782e-16 3.552177e-23 3.168356e-09  
## 1005 9.998364e-01 9.915866e-01 9.999968e-01  
## 843 1.000000e+00 9.999997e-01 1.000000e+00

There are only 6 values represented in this, which are lower, medium and bigger.

When the predicted probability of non-authenticity is 2.18e-20, with 95% confidence the mean of the probability of non-authenticity is between 1.9e-29 and 2.5e-11.

When the predicted probability of non-authenticity is 2.37e-14, with 95% confidence the mean of the probability of non-authenticity is between 1.69e-20 and 3.31e-08.

When the predicted probability of non-authenticity is approximately 1, with 95% confidence the mean of the probability of non-authenticity is between 0.992 and 1.

## Model Diagnostics

#### Likelihood Ratio Test

pchisq(fit\_int2\_aic$null.deviance - fit\_int2\_aic$deviance,  
fit\_int2\_aic$df.null - fit\_int2\_aic$df.residual,  
lower.tail = F)

## [1] 0

Our final model results in a highly significant difference between the null deviance and residual deviance as we have a p-value of 0.

#### Goodness-of-Fit Test

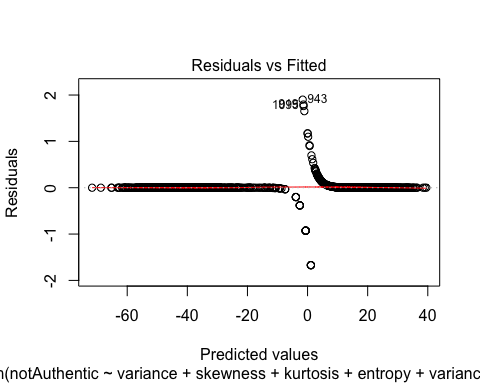
pchisq(deviance(fit\_int2\_aic), df.residual(fit\_int2\_aic), lower=F)

## [1] 1

As we do not have multiple observations for the most combinations of predictor values, the goodness-of-fit test is not the ideal test for assessing utility. However, we can see that the p-value is one which indicates a singificant difference between the deviances.

### Deviance Residuals v/s Fitted Values

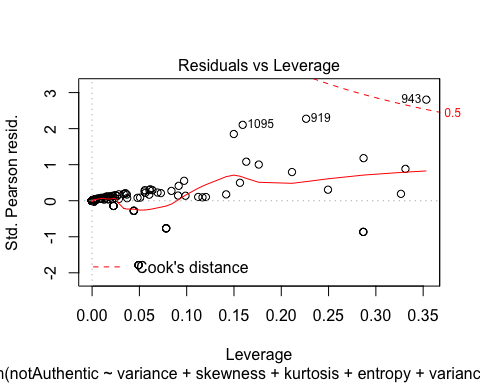
plot(fit\_int2\_aic, which = 1)



The trendline in the residual v/s fitted plot shows little to no deviation from the horizontal zero line showing that the distribution of the residuals is not concerning: it shows clear linearity. Further, there do appear to be a few points with high residual value (line observation 943).

### Influence Diagnostic Analysis

plot(fit\_int2\_aic, which = 5)



In this ‘Residuals vs Leverage plot’, we are trying to check unusual points in this dataset. On the residuals vs leverage plot, none of the residuals are greater than +/- 3 standardized pearson residuals, so we don’t appear to have any major residual outliers.

Let’s calculate very high values for Leverage in our dataset:

Using the formula (3(k+1))/n we found that our very high leverage value is 0.015. There are many data points which have high leverage points in this plot. Thus, there are multiple points with high amounts of influence.

We can also see the Cook’s distance on the plot, and we can see observation 943 is slightly above cook’s distance cutoff of 0.5. This might has high influence (leverage) and a relatively high standardizeed residual.

### ROC Plot

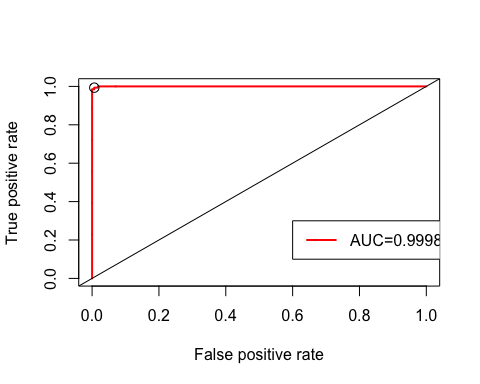
# ROC curve - install package ROCR  
par (mfrow=c(1,1))  
library(ROCR)  
pred1 <- prediction(fit\_int2\_aic$fitted.values, fit\_int2\_aic$y)  
perf1 <- performance(pred1,"tpr","fpr")  
auc1 <- performance(pred1,"auc")@y.values[[1]]  
auc1

## [1] 0.9998343

plot(perf1, lwd=2, col=2)  
abline(0,1)  
legend(0.6, 0.3, c(paste ("AUC=", round (auc1, 4), sep="")), lwd=2, col=2)  
  
# Extract the X and Y values from the ROC plot, as well as the probability cutoffs  
roc.x = slot (perf1, "x.values") [[1]]  
roc.y = slot (perf1, "y.values") [[1]]  
cutoffs = slot (perf1, "alpha.values") [[1]]  
  
auc.table = cbind.data.frame(cutoff=pred1@cutoffs,   
 tp=pred1@tp, fp=pred1@fp, tn=pred1@tn, fn=pred1@fn)  
names (auc.table) = c("Cutoff", "TP", "FP", "TN", "FN")  
auc.table$sensitivity = auc.table$TP / (auc.table$TP + auc.table$FN)  
auc.table$specificity = auc.table$TN / (auc.table$TN + auc.table$FP)  
auc.table$FalsePosRate = 1 - auc.table$specificity  
auc.table$sens\_spec = auc.table$sensitivity + auc.table$specificity  
  
# Find the row(s) in the AUC table where sensitivity + specificity is maximized  
  
auc.best = auc.table [auc.table$sens\_spec == max (auc.table$sens\_spec),]  
auc.best

## Cutoff TP FP TN FN sensitivity specificity FalsePosRate sens\_spec  
## 1102 0.501428 606 5 757 4 0.9934426 0.9934383 0.00656168 1.986881

# Plot the maximum point(s) on the ROC plot  
  
points (auc.best$FalsePosRate, auc.best$sensitivity, cex=1.3)



The ROC curve shows the trade-off between sensitivity and specificity. We can see our model does a great job because we have an AUC of 0.9998, showing that the model correctly predicted 99.98% of cases (nearly all the cases).

We can also see that our model does a good job at predicting both non-failures and failures.

We can see that the specifity, which is the probability of true negative prediction is 0.9934.

The sensitivity of the model is 0.9934, which shows us the probability of a true positive prediction.

## Conclusion

After going through all the necessary steps to make the model more reliable and predictable, the final model uses the AIC criterion of the partially-interacted effects which includes multiple variables and interactions that are significant to make a prediction. We conducted step-wise regresison on our first-order model and an interacted model and concluded that the interacted model is a better fit because of a smaller residual deviance.

The final model has a residual deviance 43.657 units (of authenticity) and the AIC score is 57.657.

The final model shows that most of the variables of the model are signficant, except entropy. However, the ROC curve and the model diagnostics show that the final model is a close, if not completely, perfect predictive model of whether a banknote is fake or authentic based on certain image chracteristics. Also, with the influence diagnostic model, we found out that there a lot of points that have a large leverage value and there is an observation that has a higher Cook’s distance value than the cutpff of 0.5. For further investigation, we would suggest looking at these points closely to further understand thesee anomolies and how to deal with them.

We would recommend seeing if there are any more features that can be recorded from an image but this model is a good, if not great, model to use to understand the authenticity of banknotes.