DTSC 560 - Logistic Regression Analysis

Importing all the necessary packages

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## #BlackLivesMatter
##
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
##     lift
##
##
## Type 'citation("pROC")' for a citation.
##
##
## Attaching package: 'pROC'
##
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
##
##
## Loaded ROSE 0.0-4
```

Loading the data

1) Generate summary statistics for the variables in the insurance.csv dataset. Quiz question #1: What percentage of customers have submitted a recent claim?

Table 1: summary statistics for the variables in the insurance

	Total		
	Overall, N=7232 (100%)[1]	**0**, N=5324 (74%)[1]	
KIDSDRIV	860 (11.9%)	528 (9.9%)	
AGE			
Mean (SD) Median (IQR)	45 (9) 45 (39, 51)	45 (8) 46 (40, 51)	
Range	16, 81	16, 81	
HOMEKIDS	2,548 (35.2%)	1,673 (31.4%)	
INCOME			
Mean (SD) Median (IQR)	61,912 (47,495) 54,009 (28,117, 86,166)	65,752 (48,397) 58,124 (30,876, 91,672)	51,196
Range	0, 367,030	0, 367,030	
HOMEOWN	4,796 (66.3%)	3,739 (70.2%)	
MSTATUS	4,324 (59.8%)	3,393 (63.7%)	
GENDER	3,880 (53.7%)	2,830 (53.2%)	
EDUCATION	4,092 (56.6%)	3,226 (60.6%)	
TRAVTIME			
Mean (SD) Median (IQR)	34 (16) 33 (23, 44)	33 (16) 32 (22, 43)	
Range	5, 142	5, 142	
CAR_USE	2,697 (37.3%)	1,760 (33.1%)	
BLUEBOOK			
Mean (SD) Median (IQR)	15,766 (8,418) 14,460 (9,360, 20,870)	16,284 (8,385) 15,060 (10,030, 21,330)	14,320
Range	1,500, 69,740	1,500, 69,740	
TWC			
Mean (SD) Median (IQR)	5.4 (4.2) 4.0 (1.0, 7.0)	5.6 (4.2) 6.0 (1.0, 8.0)	
Range	1.0, 25.0	1.0, 25.0	
RED_CAR	2,100 (29.0%)	1,557 (29.2%)	
CLM_BEF	2,784 (38.5%)	1,662 (31.2%)	
REVOKED	890 (12.3%)	496 (9.3%)	
MVR_PTS	3,984 (55.1%)	2,692 (50.6%)	
CAR_AGE			
Mean (SD) Median (IQR)	8.3 (5.7) 8.0 (1.0, 12.0)	8.7 (5.7) 9.0 (4.0, 13.0)	
Range	-3.0, 28.0	0.0, 28.0	
URBANICITY	5,769 (79.8%)	3,963 (74.4%)	

[1] n (%)

The percentage of customers have submitted a recent claim is 26%

2) Partition the dataset into a training, validation, and test set, using a 60%-20%-20% split.
Quiz question #2: How many observations are in the test set?

```
## Rows: 1,447
## Columns: 19
## $ CLAIM      <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0,~
## $ KIDSDRIV   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ AGE        <dbl> 50, 45, 39, 31, 33, 52, 33, 63, 49, 44, 54, 55, 33, 41, 37,~
## $ HOMEKIDS   <dbl> 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0,~
## $ INCOME     <dbl> 114986, 0, 51884, 18903, 79606, 51278, 20822, 74426, 0, 139~
## $ HOMEOWN    <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1,~
## $ MSTATUS    <dbl> 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0,~
## $ GENDER     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1,~
## $ EDUCATION  <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1,~
## $ TRAVTIME   <dbl> 36, 48, 43, 48, 27, 37, 46, 11, 42, 35, 61, 5, 14, 34, 29, ~
## $ CAR_USE    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0,~
## $ BLUEBOOK   <dbl> 18000, 6000, 10590, 17330, 16830, 1500, 6880, 27600, 6400, ~
## $ TWC        <dbl> 1, 1, 6, 4, 9, 4, 13, 1, 10, 10, 6, 6, 1, 3, 1, 4, 1, 1, 4,~
## $ RED_CAR    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0,~
## $ CLM_BEF    <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,~
## $ REVOKED    <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ MVR_PTS    <dbl> 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1,~
## $ CAR_AGE    <dbl> 17, 5, 13, 7, 15, 10, 1, 18, 14, 1, 11, 13, 5, 18, 10, 11, ~
## $ URBANICITY <dbl> 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1,~
```

3) We don't have a severe class imbalance in the insurance dataset, so we're going to start with fitting a model to the training set. Conduct a logistic regression analysis using the training data frame with CLAIM as the outcome variable and all the other variables in the dataset as predictor variables.

```
##
## Call:
## glm(formula = CLAIM ~ ., family = "binomial", data = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.798e+00  3.439e-01  -8.136 4.09e-16 ***
## KIDSDRIV     6.928e-01  1.330e-01   5.210 1.89e-07 ***
## AGE         -2.906e-03  5.532e-03  -0.525  0.59942
## HOMEKIDS     3.226e-01  1.108e-01   2.912  0.00359 **
## INCOME      -5.172e-06  1.124e-06  -4.601 4.20e-06 ***
## HOMEOWN     -2.619e-01  9.693e-02  -2.702  0.00690 **
## MSTATUS     -7.306e-01  9.690e-02  -7.540 4.69e-14 ***
## GENDER       1.581e-01  1.098e-01   1.439  0.15009
## EDUCATION   -5.170e-01  1.113e-01  -4.647 3.36e-06 ***
## TRAVTIME     1.659e-02  2.582e-03   6.426 1.31e-10 ***
## CAR_USE      9.752e-01  8.747e-02  11.148  < 2e-16 ***
## BLUEBOOK    -3.224e-05  5.477e-06  -5.887 3.93e-09 ***
## TWC         -5.240e-02  1.005e-02  -5.216 1.83e-07 ***
## RED_CAR     -1.193e-01  1.165e-01  -1.024  0.30591
## CLM_BEF      6.248e-01  8.365e-02   7.469 8.07e-14 ***
## REVOKED      6.592e-01  1.095e-01   6.019 1.75e-09 ***
## MVR_PTS      4.006e-01  8.609e-02   4.653 3.27e-06 ***
## CAR_AGE     -8.217e-05  9.252e-03  -0.009  0.99291
## URBANICITY   2.257e+00  1.561e-01  14.454  < 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: 5007.9  on 4338  degrees of freedom
## Residual deviance: 3914.9  on 4320  degrees of freedom
## AIC: 3952.9
##
## Number of Fisher Scoring iterations: 5
```

```
## The odd ratio:
```

```
##             odds_ratios
## (Intercept)  0.06095287
## KIDSDRIV     1.99926456
## AGE          0.99709858
## HOMEKIDS     1.38073472
## INCOME       0.99999483
## HOMEOWN      0.76960179
## MSTATUS      0.48161757
## GENDER       1.17124636
## EDUCATION    0.59629965
## TRAVTIME     1.01672903
## CAR_USE      2.65157745
## BLUEBOOK     0.99996776
## TWC          0.94895171
## RED_CAR      0.88754976
## CLM_BEF      1.86780633
## REVOKED      1.93332042
## MVR_PTS      1.49272520
## CAR_AGE      0.99991784
## URBANICITY   9.55088633
```

Quiz question #3: What is the coefficient for the KIDSDRIV variable?

The coefficient for the KIDSDRIV variable is 1.9993

Quiz question #4: What is the odds ratio for the URBANICITY variable?

The odds ratio for the URBANICITY variable is approximately 9.5509

Quiz question #5: How would you interpret the odds ratio for the URBANICITY variable? (MC)

The URBANICITY variable odds ratio means that a customer who lives in an urban has made a recent auto insurance claim 9.5509 times more than customers living in rural area.

4) Using the model you fitted in Step (3) and the validation data frame you created in Step (2), create a confusion matrix to assess the accuracy of the logistic regression model.

```
## Confusion Matrix and Statistics
##
##
## predicted_classes   0   1
##                 0 984 255
##                 1  81 126
##
##                Accuracy : 0.7676
##                  95% CI : (0.745, 0.7892)
##     No Information Rate : 0.7365
##     P-Value [Acc > NIR] : 0.003605
##
##                   Kappa : 0.2984
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9239
##             Specificity : 0.3307
##          Pos Pred Value : 0.7942
##          Neg Pred Value : 0.6087
##              Prevalence : 0.7365
##          Detection Rate : 0.6805
##    Detection Prevalence : 0.8568
##       Balanced Accuracy : 0.6273
##
##        'Positive' Class : 0
##
```

Quiz question #6: How many insurance claims (positives) did the model predict correctly?

The model predicted correctly 984 insurance claims.

Quiz question #7: What is the accuracy rate? The accuracy rate is 0.7676

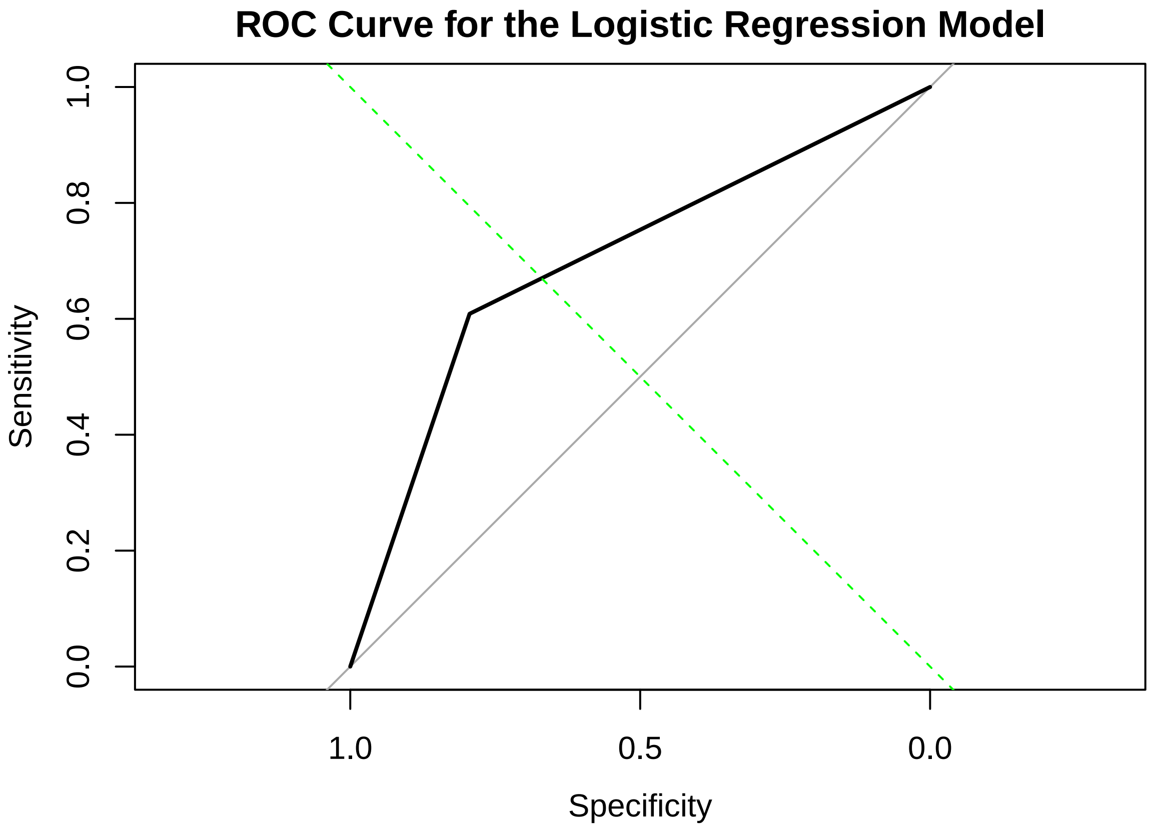
Quiz question #8: What is the sensitivity? The sensitivity is 0.9239

Quiz question #9: How would you interpret the sensitivity? (MC) Sensitivity obtained in the above question tells us that has 92.29% ability to designate an individual with a recent auto insurance claim as positive.

5) Again using the model you fitted in Step (3) and the validation data frame, create an ROC curve plot and calculate the AUC.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
## AUC: 0.7014423
```

Quiz question #10: What is the AUC?

AUC = 0.7014423

6) Even though we do not have a severe class imbalance in our data, let’s try addressing our moderate class imbalance to see if it improves our model accuracy. Using the training set you generated in Step (2), create a new training subset using the oversampling method.

```
##
##    0    1
## 3194 3194
```

Quiz question #11: In this new training subset generated from oversampling, how many observations are in the class that has made a recent auto claim (“Yes”)?

3194 observations are in the class that has made a recent auto claim (“Yes”)

7) Conduct a logistic regression analysis using the new oversampled training subset with CLAIM as the outcome variable and all the other variables in the dataset as predictor variables.

```
##
## Call:
## glm(formula = CLAIM ~ ., family = "binomial", data = data_balanced_over)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.158e+00  2.476e-01  -8.714  < 2e-16 ***
## KIDSDRIV     6.897e-01  1.051e-01   6.561 5.36e-11 ***
## AGE          2.462e-03  3.994e-03   0.616  0.53768
## HOMEKIDS     3.899e-01  8.414e-02   4.635 3.58e-06 ***
## INCOME      -5.476e-06  8.189e-07  -6.687 2.28e-11 ***
## HOMEOWN     -2.355e-01  7.484e-02  -3.147  0.00165 **
## MSTATUS     -7.643e-01  7.477e-02 -10.223  < 2e-16 ***
## GENDER       2.109e-01  8.393e-02   2.513  0.01197 *
## EDUCATION   -6.461e-01  8.579e-02  -7.532 4.99e-14 ***
## TRAVTIME     1.640e-02  1.973e-03   8.313  < 2e-16 ***
## CAR_USE      1.043e+00  6.905e-02  15.111  < 2e-16 ***
## BLUEBOOK    -3.525e-05  4.122e-06  -8.551  < 2e-16 ***
## TWC         -5.644e-02  7.523e-03  -7.502 6.28e-14 ***
## RED_CAR     -3.613e-02  8.799e-02  -0.411  0.68135
## CLM_BEF      6.884e-01  6.308e-02  10.912  < 2e-16 ***
## REVOKED      6.981e-01  8.634e-02   8.085 6.20e-16 ***
## MVR_PTS      4.705e-01  6.420e-02   7.328 2.33e-13 ***
## CAR_AGE      1.015e-02  6.968e-03   1.456  0.14543
## URBANICITY   2.275e+00  1.070e-01  21.269  < 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: 8855.6  on 6387  degrees of freedom
## Residual deviance: 6669.4  on 6369  degrees of freedom
## AIC: 6707.4
##
## Number of Fisher Scoring iterations: 4
```

8) Using the model you fitted in Step (7) and the validation data frame you created in Step (2), create a confusion matrix to assess the accuracy of the logistic regression model.

```
## Confusion Matrix and Statistics
##
##
## predicted_classes1   0   1
##                  0 774 119
##                  1 291 262
##
##                Accuracy : 0.7165
##                  95% CI : (0.6925, 0.7396)
##     No Information Rate : 0.7365
##     P-Value [Acc > NIR] : 0.9601
##
##                   Kappa : 0.362
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7268
##             Specificity : 0.6877
##          Pos Pred Value : 0.8667
##          Neg Pred Value : 0.4738
##              Prevalence : 0.7365
##          Detection Rate : 0.5353
##    Detection Prevalence : 0.6176
##       Balanced Accuracy : 0.7072
##
##        'Positive' Class : 0
##
```

Quiz question #12: What is the accuracy rate?

Accuracy = 0.7158

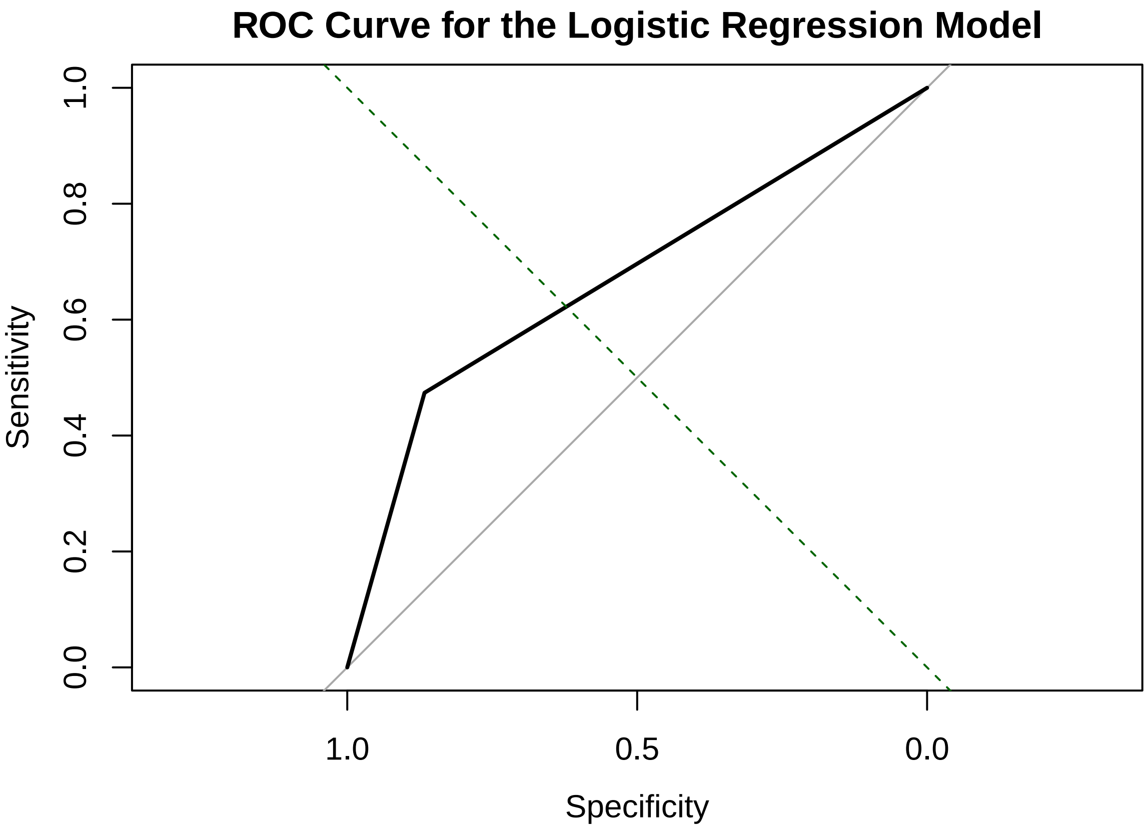
Quiz question #13: What is the sensitivity?

Sensitivity = 0.7202

9) Again using the model you fitted in Step (7) and the validation data frame, create an ROC curve plot and calculate the AUC.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
## AUC: 0.6702604
```

Quiz question #14: What is the AUC?

AUC = 0.6725446

Quiz question #15: What do you notice about this AUC value as compared to the AUC value for the previous model? (MC)

AUC value in the second model is smaller the AUC value for the first model (0.6725 < 0.7014).

10) Let's say that for this insurance company, sensitivity is more important than overall accuracy and the cost of false positives is lower than the cost of false negatives, so we will use the logistic regression model fitted to the over sampled training subset. Using the model generated in Step (7) and the test set you created in Step (2), create a confusion matrix to assess the accuracy of the logistic regression model on the test data frame.

```
## Confusion Matrix and Statistics
##
##
## predicted_classes2   0   1
##                  0 760 100
##                  1 305 282
##
##                Accuracy : 0.7201
##                  95% CI : (0.6962, 0.7431)
##     No Information Rate : 0.736
##     P-Value [Acc > NIR] : 0.9188
##
##                   Kappa : 0.3855
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7136
##             Specificity : 0.7382
##          Pos Pred Value : 0.8837
##          Neg Pred Value : 0.4804
##              Prevalence : 0.7360
##          Detection Rate : 0.5252
##    Detection Prevalence : 0.5943
##       Balanced Accuracy : 0.7259
##
##        'Positive' Class : 0
##
```

Quiz question #16: How many insurance claims (positives) did the model predict correctly using the test set?

The model predicted correctly 749 insurance claims (positives) using the test set.

Quiz question #17: What is the accuracy rate?

Accuracy = 0.7132

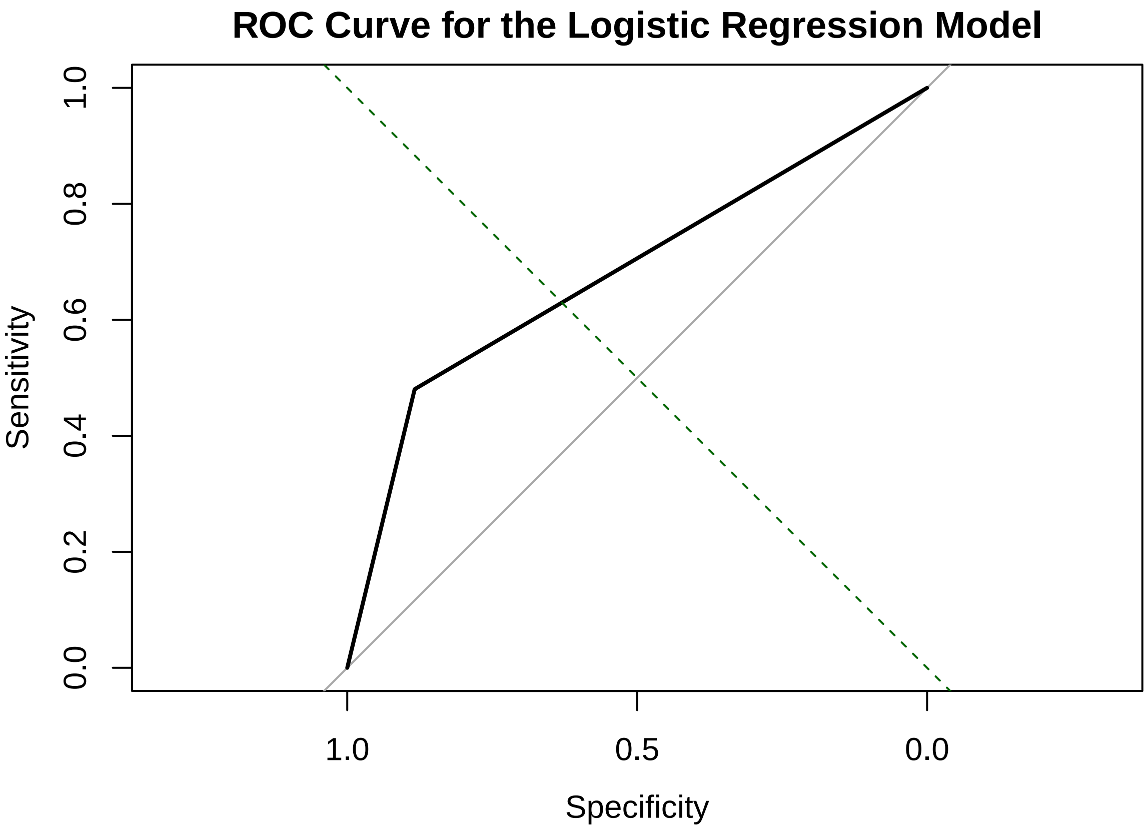
Quiz question #18: What is the sensitivity?

Sensitivity = 0.7033

11) Again using the model you fitted in Step (7) and the test data frame, create an ROC curve plot and calculate the AUC.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
## AUC: 0.6820649
```

Quiz question #19: What is the AUC?

AUC = 0.6779

12) Now we'll use the model fitted to the oversampled training subset to make predictions about whether new customers will make auto insurance claims. Using the data contained in the csv file "insurance_predictions.csv", predict the probability scores for insurance claims for ten new customers.

Quiz question #20: What is the predicted probability of making an insurance claim for new customer #1?

```
##    predicted probability
## 1              0.6249027
## 2              0.4143574
## 3              0.1881750
## 4              0.1858973
## 5              0.1472866
## 6              0.7593088
## 7              0.4201541
## 8              0.6018195
## 9              0.4371518
## 10             0.3566545
```