DTSC 560
Data Science for Business

Module 5 Assignment: Logistic Regression Analysis

For this assignment, you will conduct a logistic regression analysis in R. **You will not be turning in any code or output; rather you'll do the analysis and use the output to answer questions for the associated assignment quiz on Brightspace.**

Please read these instructions carefully to be able to align your answers with the associated Module 5 assignment quiz in Brightspace.

Data: insurance.csv (download from Module 5 on Brightspace)
We are using a dataset of information from 7,232 car insurance customers, some of whom have made insurance claims and some who haven't.

You will also use a second dataset of new customers to predict the probability of new customers making insurance claims. **That dataset is called insurance_predictions.csv** (download from Module 5 on Brightspace).

Background: For this assignment, you work at an auto insurance company and you would like to predict the probability of insurance claims based on different customer characteristics. Your business question is: "What is the probability that a customer will make an auto insurance claim based on certain characteristics?"

Variables: The variables in this dataset include:

- CLAIM: Whether a customer has made a recent auto insurance claim (No = 0, Yes = 1)
- KIDSDRIV: Whether a customer has children that are driving (No = 0, Yes = 1) ● AGE: Age of driver in years
- HOMEKIDS: Whether a customer has children at home (No = 0, Yes = 1)
- INCOME: Income in dollars

- HOMEOWN: Whether a customer owns a home (No = 0, Yes = 1)
- MSTATUS: Whether married (No = 0, Yes = 1)
- GENDER: Gender (Male = 0, Female = 1)
- EDUCATION: Level of education (High School only = 0, College or beyond = 1)
- TRAVTIME: Commute time to work in minutes
- CAR_USE: Type of car use, private or commercial (Private = 0, Commercial = 1)
- BLUEBOOK: Value of vehicle in dollars
- TWC: Customer time with insurance company in years
- RED_CAR: Whether a customer's car is red (No = 0, Yes = 1)
- CLM_BEF: Whether a customer has made a previous claim in the last five years (No = 0, Yes = 1)
- REVOKED: Whether a customer has had their license revoked (No = 0, Yes = 1) ● MVR_PTS: Whether a customer has motor vehicle record points (traffic tickets) (No = 0, Yes = 1)
- CAR_AGE: Vehicle age in years
- URBANICITY: Whether a customer lives in an urban or rural area (Rural = 0, Urban = 1)

Assignment Steps:

Carry out the steps below to complete the assignment, then answer the questions in the Module 5 Assignment Quiz on Brightspace. The quiz questions are included here, with their numbers, if you prefer to answer them as you are doing the assignment and enter them in the Brightspace quiz all at once (multiple choice questions are labeled "MC").

Step 1) Generate summary statistics for the variables in the insurance.csv dataset.

Quiz question #1: What percentage of customers have submitted a recent claim?

Step 2) Partition the dataset into a training, validation, and test set, using a 60%-20%-20% split.

****IMPORTANT: In order to get results that align with the correct answers in the assignment quiz, when you are partitioning your dataset you MUST set the seed value to 42 using the set.seed () function. If you do not do this, you will not be able to reproduce the answers that correspond with the assignment quiz.**

Quiz question #2: How many observations are in the test set?

Step 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.

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

Quiz question #4: What is the odds ratio for the URBANICITY variable?
Quiz question #5: How would you interpret the odds ratio for the URBANICITY variable? (MC)

Step 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.

Quiz question #6: How many insurance claims (positives) did the model predict correctly?
Quiz question #7: What is the accuracy rate?
Quiz question #8: What is the sensitivity?
Quiz question #9: How would you interpret the sensitivity? (MC)

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

Quiz question #10: What is the AUC?

Step 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.

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")?

Step 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.

Step 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.

Quiz question #12: What is the accuracy rate?
Quiz question #13: What is the sensitivity?

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

Quiz question #14: What is the AUC?
Quiz question #15: What do you notice about this AUC value as compared to the AUC

value for the previous model? (MC)

Step 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 oversampled 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.
Quiz question #16: How many insurance claims (positives) did the model predict correctly using the test set?
Quiz question #17: What is the accuracy rate?
Quiz question #18: What is the sensitivity?

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

Quiz question #19: What is the AUC?

Step 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?