STOR 538 Sports Analytics

Practice 2

Objectives:

- Linear Regression
- Logistic Regression
- Model Testing Using a Test Set and Cross Validation

Introduction:

I fully expect you to complete this entire assignment by yourself without the help of another person. The work you submit should be your own, but you can use the internet to help you find solutions. You will be required to either show all your code and/or provide step-by-step instructions on how you completed the various tasks. If there is evidence that you either submitted someone else's work or failed to submit work that illustrates how you achieved certain output, expect your instructor to pursue an Honor Code violation.

For this assignment, you will be building linear regression models and logistic regression models. After showing the output from those models, you will be testing those models using a test set and cross validation. Do this assignment in the order I ask. Provide comments and section headings that clearly tell the grader what each piece of code and output is doing and what it is for.

Your analysis is based on a dataset I found on a github from user "DanielEliezer". The dataset contains a row for 37,023 extra points and field goals in the NFL. I named the dataset **kicking** and it is a CSV file.

Now, read the next section to find all of the tasks I want you to complete for this dataset.

Assignment:

- Linear Regression
    a. Start by creating a new variable called spread which is calculated based off the *home_score_pre -visiting_score_pre*
    b. Create a **histogram** or **boxplot** of the *spread* variable
    c. Then fit a linear regression model, using all data, to predict *spread* based off knowing only the home team, the away team, the quarter, and the yard line.
    d. Show only the following output from the previous linear regression. This output doesn't have to be shown in one table but can be if you would like. Alternately, it can be shown in multiple pieces of output in the order listed below. Please try to limit how many extra things you show in output that were not requested here.
        i. Coefficients and P-values from individual t-tests on Coefficients (Show in one table and this will be a large table because of the categorical variables. **It is okay if the output you show is default output that contains other things like test statistics and standard errors**)
        ii. R-Squared
        iii. Adjusted R-Squared
    e. Create a new model, using all the data, that adds the interaction between the home team and the away team to the previous model and only show the following output.
        i. R-Squared
        ii. Adjusted R-Squared
    f. Randomly split the data into 80% Training data and 20% Testing data. Refit both of the models above on the Training data and predict the spread on the testing data. Then, calculate the **RMSE** and **MAD** of both models using the differences between the predicted spread and actual spread in the test set. The grader should see how you split the data, how you fit the models on the training data, how you obtained predictions on the testing data, and how you calculated the RMSE and MAD on the test set. Create a table like seen below that contains the RMSE and MAD to organize your output.

Model	RMSE	MAD
Without Interaction	# from Calculation	# from Calculation
With Interaction	# from Calculation	# from Calculation

- Logistic Regression
    a. Subset the data for only the field goals. You will need the spread variable from the previous part. (No extra points). **Also, only keep kickers in your data who have more than 100 field goals. When you isolate the field goals and remove kickers with 100 or less field goals, you should have 14,052 observations.**
    b. The variable scored is binary where 1 indicates success and 0 indicates failure. Fit a logistic regression model, using all the data, to predict the probability of success given the yard line, the quarter, the kicker name, and the point spread. **Show a table that contains output from logistic regression that at minimum contains coefficients and p-values.**
    c. Print out a contingency table/confusion matrix from the previous model that shows the number of true positives, true negatives, false positives, and false negatives. Rough example seen below:

	Model Predicted Success	Model Predicted Failure
Field Goal Successful	# of True Positives	# Number of False Negatives
Field Goal Failed	# Number of False Positives	# Number of True Negatives

    Your table can have different labels, but the grader should know exactly what is "actual" and what is "predicted".
    d. Evaluate the logistic regression above using 10-Fold Cross Validation. **You can use a package/library to perform 10-Fold CV if you would like, but I would recommend using the steps below with a loop**. **If using a package/library, make sure you are specifically performing 10-Fold CV.**
        i. Create an empty column in the data to save your predictions.
        ii. Create a variable in the table called *fold* where each value is an integer from 1 to 10 randomly sampled. Each row should now be assigned randomly to 1 of 10 folds and each fold should contain **approximately** the same number of observations since the numbers 1 to 10 were chosen at random with equal probability.
        iii. Fit the model to the entire dataset where *fold does not equal 1*
        iv. Predict 0 or 1 using the model for all the data where *fold equals 1*
        v. Save your predicted 0's or 1's into the empty column for only the rows in the dataset where *fold equals 1*
        vi. Then, fit the model to the entire dataset where *fold does not equal 2*
        vii. Predict 0 or 1 using the model for all the data where *fold equals 2*
        viii. Save your predicted 0's or 1's into the empty column for only the rows in the dataset where *fold equals 2*

ix. **Once you know your code works for the first two folds,** repeat the pattern for all 10 folds. I recommend using a loop through the numbers 1 to 10 which are the folds.
x. After you loop through each fold and complete the pattern above for each fold, your originally "empty" column of predictions should contain no empty values, but should contain a predicted 0 or 1 for each observation.

e. Finally, print out a contingency table/confusion matrix (see part c), comparing the actual 0's and 1's to predicted 0's and 1's acquired through LOOCV.

If you plan on using R, Python, or another programming language, you need to submit a PDF file containing all of your code and output. Also, provide comments letting the grader know what each line or block of code is doing.

If you plan on doing this without using R or Python, you should still submit a PDF showing screenshots of what you did and how you did every step, showing screenshots of the exact output requested, along with a detailed explanation of every screenshot provided.

Submission:

You will need to submit one PDF file, named "regression.pdf", that contains all of your code and output that accomplishes everything I asked in the order in which it was requested. Use comments or headings that signify what code and output corresponds to the different parts requested of you above.

Rubric:

The table below shows how many points are allocated to each item along with some explanation or reminder of key points. You will lose points for not following the detailed instructions given above.

Criteria	**Points**
Linear Regression: Calculated spread and showed histogram or boxplot of spread	3 Points (*Working code with a lot of comments or detailed explanation*)
Linear Regression: First Linear Regression and showing output for all things requested.	5 Points (*Working code with a lot of comments or detailed explanation*)
Linear Regression: Second Linear Regression and showing R-sqared and Adjusted R-squared	2 Points (*Working code with a lot of comments or detailed explanation*)
Linear Regression: Split data up, fit models to train, predicted models on test, then calculated RMSE + MAD	5 Points (*Working code with a lot of comments or detailed explanation*)
Linear Regression: Created table to summarize RMSE and MAD for models	2 Points *(Follow instructions to create similar table as shown in example)*
Logistic Regression: Removed extra points and removed kickers that only appear once.	*3 Points (Showed code or process to clean the data for these issues)*
Logistic Regression: Fit Logistic Regression Model on Field Goals Only and Show Output with Coefficients and P-values	2 Points (*Working code with a lot of comments or detailed explanation*)
Logistic Regression: Create contingency table and confusion matrix showing Counts	4 Points (*1 Point per Cell in Matrix*)
Logistic Regression: Performed 10-Fold CV	3 Points (*Working code with a lot of comments or detailed explanation*)
Logistic Regression: Create contingency table and confusion matrix showing Counts on Predictions after 10-Fold CV	4 Points (*1 Point per Cell in Matrix*)
Instructions: Did everything in the exact order requested	2 Points (*Followed instructions and did the work in the order requested)*
Instructions: Only the output requested is shown. Any additional output that is unnecessary and leads to a difficulty in grading will result in loss of points.	2 Points (*This is at the discretion of the grader)*