NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PROJECT 1: MONEYBALL OLS REGRESSION
PREDICT 411

This data set contains approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. You are to use OLS ("Linear") Regression and the given statistics to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided).

DELIVERABLES:

- Your write up in PDF Format. Your write up should have four sections. Each one is described below. **(160 Points)**

- A stand-alone R program with accompanying protocol directions that can be used to score new data as it becomes available. This should include all of the data cleaning, transformations, and the regression equation(s) from your analysis. **(40 Points)**

- An EXCEL file that contains the scored records values from MONEYBALL_TEST. There will be only two columns in this file: INDEX and P_TARGET_WINS. You will be graded on how well your model performs versus my model, other professor selected benchmark models, and those of other students in the class. Be sure you have predicted values for ALL records in the MONEYBALL_TEST file. **(50 Points)**

Double check:

- Submission file in PDF format?
- Do you have a GOOD Introduction?
- Do you have a GOOD Conclusion?
- Did you write about and comment on EVERY table and graph you included in your write-up? If not, get rid of it or write something stellar! Don't data dump me – I don't like that!
- Can your write-up be read and understood by your boss at work?

WRITE UP (160 POINTS):

1. DATA EXPLORATION (40 points)

Describe the size and the variables in the MONEYBALL data set so that a manager can understand it. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

2. DATA PREPARATION (40 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value or use a decision tree)
b. Create flags to suggest if a variable was missing.
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS (40 Points)

Build at least three different LINEAR REGRESSION using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the model, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS (40 Points)

Decide on the criteria for selecting the "Best Model". Will you use a metric such as Adjusted R-Square or AIC? Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

STAND ALONE SCORING PROGRAM (40 POINTS)

Write a Stand Alone R program with accompanying protocol that will score new data and predict the number of wins. I'm leaving this open and flexible for you, but it must be something sufficiently detailed that a knowledgeable person can follow your directions and code, in order to score data independently from you. In the past, this was a stand alone program. I'm allowing you to "clean" data using EXCEL or some other venue, but you have to be clear exactly how to clean and prepare the data.

The variable with the Predicted number of Wins should be named:

P_TARGET_WINS

The R program and protocol will need to include:

a. All the variable transformations such as how to fix missing values
b. The regression formula(s)

SCORED DATA FILE (50 POINTS)

Use the R program and protocol that you wrote in the previous section to score the data file MONEYBALL_TEST. Create a file that has only TWO variables for each record:

INDEX
P_TARGET_WINS

The first variable, INDEX, will allow me to match my grading key to your predicted value. If I cannot do this, you won't get a grade. So please include this value. The second value, P_TARGET_WINS is the number of wins you believe the team will have in season based upon the data given to you. ***Be sure you have a predicted value for every record!***

Your values will be compared against …

- A Perfect Model
- Instructor's Model
- Performance of Other Students
- Predict the Average value for everybody (MEAN)
- Random Model

If your model is not better than simply using an AVERAGE Y value, your grade will suffer
If your model is not better than generating a RANDOM value, your grade will suffer a lot
If you model beats mine, you can teach this class next quarter! Just kidding! I need the money!