NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PROJECT 1: MONEYBALL OLS REGRESSION
MSDS (Predict) 411

INTRODUCTION

This project also builds on your OLS Regression knowledge from 401 and 410, and hopefully expands it considerably. Here, you will be using the MONEYBALL data set. 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 a given year, with all of the statistics adjusted to match the performance of a 162 game season. In general, you are to use OLS ("Linear") Regression and the given data/statistics to predict the Y-Response Variable which is the number of wins (TARGET_WINS) for the team. You can only use the variables given to you (or variables that you derive from the variables provided). The kicker for this project, is that the models you develop to predict the number of wins for the team must also be exported – or generalized – and applied to data from which the model(s) was not fit. This is the "proof in the pudding" for predictive modeling – can you make accurate prediction?

To this end, this project is both a teaching tool and an assessment. Because this project is a teaching tool, the project is divided into specific tasks. The tasks lead you through different kinds of analyses. These analyses are designed to provide you some perspective and insight into aspects of modeling that you may not have covered in 401 or 410. As in all knowledge derived from statistical methods/models, it is expected that ***what*** you will learn about modeling will, at least in part, be inferred from the statistical results. Pay attention and be looking for lessons learned as you go.

I ask, and strongly advise, that you complete the tasks in order. Do not jump ahead – the learning rationale is sequential. Be linear – do the tasks in order – do exactly what the tasks ask you to do. There will be some flexibility within each task, so do not expect that what you come up with will be exactly the same as everyone else. There are steps in each task that ask you to document what you have done. Please do the writing for these tasks as you go along. DO NOT save the writing for the end – that will be a disaster. Your final write-up for the project should clearly have heading and sub-heading labels for each task and step within tasks. The task directions will tell you specifically what to include in the write-up.

Finally, a primary learning outcome is for you to gain experience conducting predictive modeling in open-ended contexts. As such, the last task will ask to start over from the beginning

and put together your best possible predictive model. For this open-ended modeling task, you will be graded on how well your model predicts, relative to the professor's model, your classmate's models, and some benchmark models. Is always fun to see where we stack up!

One last cautionary comment. This first project is a very intensive statistical programming project. Do not underestimate this. This project WILL TAKE considerable time, so plan accordingly. Do NOT wait until the last minute to work on this assignment. You should be working on this assignment steadily over the 3-week time span of the first unit.

Each task is worth 50 points. The task directions follow.

TASK 1: DEVELOPMENT OF BASE MODEL (50 points)

Use your OLS Regression knowledge from the 401 and 410 classes to develop a predictive model. We will call this model your Base Model. In order to do this, you will need to complete the following steps:

i. Download the MONEYBALL data dictionary. Read through it so you know something about the variables in the MONEYBALL dataset.

ii. Download the file MONEYBALL.CSV and import this data into RStudio.

iii. Conduct an Exploratory Data Analysis (EDA) of the MONEYBALL DATA and clean the data. Feel free to use ideas from your prior experiences in 410. Some suggestions for things that you could do are:

   a. Histograms for each continuous variable
   b. Means, standard deviations, minimum, maximum, median for all continuous variables
   c. Are variables correlated to the target variable (TARGET_WINS) or to other possible explanatory variables?
   d. Are any of the variables with missing values that need to be imputed or "fixed"? Fix missing values (maybe with a Mean or Median value or use a decision tree). Are there variables with so many missing values that the entire variable should be eliminated from the analysis?
   e. Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing.
   f. Do any of the variables need a mathematical transformation, such as log or square root? Create new variables with these transformations and add them to the end of the dataset.
   g. Create any new variables that you are interested in. A lot of people are seriously into modeling baseball data. I'm sure with a little bit of google searching you can find variables like slugging percentage or power pitching that you would want to compute and potentially use. This is totally voluntary

on your part. Not at all required. Do this if you have the interest or think such variables might be of value.

Please do NOT treat this as a check list of things you must do to complete the assignment. The EDA is your responsibility. You should have your own thoughts about this step based on your prior experiences with 401 and 410.

iv. Retain the logic and R-code you developed for creating new variables, imputing missing values, and handling extreme scores. These exact same computations and logic will be needed when it comes to exporting your model in Task 4. I STRONGLY suggest writing this down in a log somewhere so you do not lose it! SAVE THE R CODE!

v. Write a description of what you did in performing your EDA and data cleaning. Describe what you did and what you found so that a manager can understand it. Consider that too much detail will cause a manager to lose interest. DO NOT DATA DUMP! If you include a graph, you must describe and discuss that graph! Similarly, too little detail will make the manager consider that you aren't doing your job or that you were not careful. Your reputation is at stake! This description is to be included in your final project write-up.

vi. Given all of the variables you've retained and created, plus the spanking clean data you've scrubbed, you now need to fit an OLS Regression model to predict the response variable, TARGET_WINS (Y). What should you do? There is not one right way of approaching this problem. In 410, you should have learned about automated variable selection procedures. Many analysts and modelers just brute force the problem and use these methods. Let's be like them! For simplicity sake and a degree of uniformity of experience, please use an automated variable selection procedure like FORWARD, BACKWARD or STEPWISE Regression to obtain an OLS regression model to predict TARGET_WINS (Y). The resultant regression model will be your BASE Model.

vii. Write a description of the method you employed to decide upon and fit your OLS BASE regression model. In additions, report your BASE OLS regression model. This means you explicitly write out the mathematical expression for this model:

$$Y\_hat = 1.985 + 5.256*X1 + 4.323*X2 - 10.652*X3 + 1.211*LOG(X1) + \ldots..$$

Naturally, you will use the actual names of the variables in your write up. Then interpret each of the coefficients. Interpret all of the parameter estimates (i.e. the coefficients or betas). Also report any key goodness of fit statistics you think are essential, like Adjusted R-squared.

viii. *Deliverables – Task 1:* Your project report should have a section that is headed as TASK 1 – DEVELOPMENT OF BASE MODEL. This section of the project report should contain:

1. *Description of the EDA and data cleaning (see step v)*

2. *Description of the modeling approach, the BASE model equation, interpretation of coefficients, and goodness of fit statistics (see step vii)*

TASK 2: BASE MODEL SIMPLIFICATION AND REFINEMENT (50 points)

Most likely, your BASE Model will have 12, 15, or 20, may contain more explanatory variables in your BASE Regression model. Did this happen to you? Let's digest this a bit.

Learning Perspective

First, look at all of the coefficient estimates in your BASE regression model. Are all of these estimates logically consistent with what should happen in the real world? For example, if batting increases, does the number of predicted wins increase? That would be logical and reasonable. If ERA decreases (i.e. pitching gets better and the other team doesn't score as much), does the number of wins increase (i.e. you should have a negative coefficient here). Are any of these parameter estimates logically the wrong way? I think that's a problem, do you? Why or why not?

Second, notice the hypothesis tests for all the parameter estimates. Are some of these tests not statistically significant? That may be an issue. This can happen from an automated variable selection procedure. Similarly, notice the parameter estimates. Are some of them close to zero? Will such a variable contribute much to a predicted outcome if the coefficient is zero?

Third, note the size of this data base. How many records are there? When you have large databases like this one, you have A LOT of statistical power. Remember that idea from 401? Now, recall how automated variable selection procedures work. How do they work? Right! They use the concept of statistical significance and hypothesis tests to decide upon which variables to retain in the model. The trouble is, that when you have a TON of statistical power, hypothesis tests will reject the null hypothesis (beta=0) even if there is only a minimal change to the parameter estimate. In other words, you can have a statistically significant test for a beta but that variable may have little or almost no predictive value. Over the huge number of hypothesis tests conducted in an automated variable selection application, you will get variables included in the final model that should not be there. They look to be statistically significant, but really are not PREDICTIVE. In such a situation, you may have an over specified model – too many variables. Further, these procedures tend to over capitalize on the idiosyncrasies of the data in hand. Usually, such models DO NOT EXPORT WELL! In other words, there may be a big difference between the TRUTH model and your BASE predictive model. We will check on this in Task 4 as best we can.

The upshot here, and the lesson to be learned, is that the result of an automated variable selection procedure is a place to START the modeling process, not END the modeling process. At best, it gives you a good set of variables to work with. Ideally, you should want as simple a model as

possible, that still is sufficiently predictive. Parsimony! Is a wonderful word and important concept for modeling purposes! What this means is that you DO NOT have to have a MAX R-Squared model in order to be successful EXPORTING your predictive model.

Modeling Steps

For Task 2, we will start with your BASE Regression model and prune it back a bit. In doing this, we will look at Adjusted R-squared change.

Please complete the following steps for Task 2:

i. Any explanatory variable (X) that is NOT in your BASE Regression model is excluded from contention. Consider the set of explanatory (X) variables in your BASE Regression model to be the only possible variables that can be included in your upcoming models. You may want to create a new dataframe with only the X variables in contention and TARGET_WINS(Y). This is optional.

ii. Obtain the correlations of all of the explanatory variables in your BASE Regression model with TARGET_WINS(Y). Consider ranking or ordering the explanatory variables by their correlation with Y. It may be helpful. Ranking or ordering is optional.

A Side Discussion – For your knowledge and learning

Based on these correlations and the coefficients/test statistics from the BASE Regression model results from Task 1, you can start to identify variables that may not be adding much to the predictive value of your BASE Regression model. You can isolate and estimate the predictive value of a variable by calculating Change in Adjusted R-squared.

For example, suppose you have 20 explanatory variables in your BASE Regression Model and X20 is in contention to be removed because you feel it does not really add explanatory value. Fit a full model to predict Y using all 20 explanatory variables and note the Adjusted R-squared value. Then fit a reduced model to predict Y using 19 explanatory variables when you excluded X20. Note the Adjusted R-squared value for the 19 variable model. The difference between the two Adjusted R-squared values is the change in Adjusted R-squared due only to X20. If this change in Adjusted R-squared is minimal, get rid of X20. If X20 contributes substantively, then keep it.

You could then simply repeat such a process to remove variables that are not really predictive. In my practice, what I usually see, though not always, is that you end up with somewhere between 1 and 6 or 7 variables in the model that are sufficiently predictive. The others just don't help, in a predictive sense, even if they are statistically significant. You may ask, "what is the cut off value for inclusion versus exclusion?" The answer is, I don't know for sure. No one does. Pick the value that makes the most sense and feels best to you. You're the one who will have to explain this to your boss! But, as a guide, I always keep in mind what Adjusted R-squared measures – namely, the proportion of variability in Y that is explained by the regression

model. If a variable explains barely a fraction of a percent of this variability in Y, do you really want that variable in your model? That's your justification. What is good enough, is up to you!

Now, back to the modeling steps!

iii. Parse through the explanatory variables in your BASE regression model, removing those variables that you find are not sufficiently predictive. Do this until you've decided on which variables to retain in the model. I anticipate that you may have between 1 and 7 variables in this model, but I don't know for sure. You need to find out. Once you are done parsing, call the resultant model your PARSIMONY model.

iv. Fit your PARSIMONY model. Report the coefficient estimates, coefficient interpretations, and goodness of fit statistics. We will compare how well this model predicts relative to your BASE model in Task 4.

v. *Deliverables – Task 2:* Your project report should have a section that is headed as TASK 2 – The Parsimony Model. This section of the project report should contain:

1. *Description of the procedure you used for obtaining the PARSIMONY model. This should include any criteria you used.*

2. *Report the PARSIMONY model, interpret the parameter estimates (i.e. the betas), and report the goodness of fit statistics (see step iv).*

TASK 3: CONTEXTUAL MODELING APPROACH (50 points)

Data doesn't just rise up in a vacuum. It comes from some data collection process and, as such, has a pedigree. We haven't really paid any attention to the quality or context of the data, yet.

Learning Perspective

Some statisticians, analysts and modelers don't use context at all. It is all about the numbers and the descriptive statistics. They go for the technical fixes, for example trimmed means. Just drop the top and bottom 10% or records as being too extreme, model the remainder, and all will be good. But, my experience as a social science statistician, you run the risk of throwing away legitimate evidence that may be highly valuable. Since all data does arise in context, paying attention to that context can dramatically help in modeling and interpreting the final result.

You've probably noticed that in our MONEYBALL dataset, there is very little information about context. There isn't a YEAR variable or a TEAM variable. We don't know if records are legitimate or not. About all we know, is the data comes from baseball teams between the years 1871 to 2006. Time to put on your data detective cap! What do you know about baseball during this time frame? Has baseball always been the same? In what ways has it been changing? Could historical changes be observed in, or influence, the data? I don't know, but

you could do some internet searching and reading to gain more background knowledge about the game of baseball and, hence, this data. I routinely do such things when I have a dataset to analyze and I don't know the field well.

Now, some basic statistical questions: a) what is the population of interest? b) is there only 1 population or are there many different populations? c) why would one think the relationships between variables would remain constant over the entire span of time from which the data was obtained? I hope you get my point. If relationships many not reasonably be consistent over time, you may want to find ways to group data and then have different predictive models for the different groups. How you answer such questions will lead you down very different analytical paths. I'll leave this to you for Task 5.

For now, let's start simple and use the data. If you look at the variable hits allowed, some of the records have over 5,000 hits allowed. Well, in a 162 game season, that translates into about 30+ hits per game on average. That just doesn't happen in professional baseball today. Even if you scale back to 2500 hits allowed, this translates into about 15/game, which is still twice that of the current major league average and outside any band of variability one might construct from modern era baseball data (YES – I did go on line to find statistics to check on this! It is what data detectives do!). Most of the variables in the MONEYBALL dataset have this same kind of issue. So I ask myself, "Could such data points come into existence and still be legitimate?" I understand the data runs from 1871 to 2006. The modern era of baseball didn't start until around 1967 or so, and in the early days, batters would tell pitchers where to place the pitch so that they could hit it. It was a gentleman's game after all. The point was to have fun and hit the ball! So, those extreme scores could come from a NON-MODERN era time of baseball. Simple transformations of variables, like from total hits to hits per game, can give contextual insight into the nature of the data if you combine the transformation with a little bit of easily obtainable knowledge from the internet. This points to looking at Modern Era records versus Non-Modern era records.

You might ask, "How can I use the Contextual Information to help my modeling effort?" What I do is create and use FLAGS, or indicator variables. Through the use of FLAGS, you could identify those records that pass a "smell test" – i.e. data that is consistent with modern era baseball statistics. Records that don't or that are too far out of the range of "normal" modern professional baseball get a FLAG (value of 1). You don't need to delete or change the records or the data. If data is missing, you can set a MISSINGFLAG to 1. Further, you can set a separate FLAG for each explanatory variable. Data gets seriously modeled, only if it doesn't smell bad. In other words, records that do not have any flags across all the variables in the dataset. You can call such records, consistent with modern era baseball. Such records get set aside and modeled using basic OLS Regression procedures. The records that are off or FLAGGED can be modeled with a completely separate model. You can use a regression model for these records, or, you can simply use the mean of the Y's as your predicted value, or mean of the Y's for the flagged records. The point is, you can use more than 1 regression model when modeling a dataset.

The analogy is this. If you have housing data for a neighborhood where all the houses are essentially ranch style homes. But, there are a couple mansions in the neighborhood too. Why not model the ranch style homes separately from the mansions? Aren't they really from

different populations, even if they are all in the same classification of homes? In a sense you would be identifying more than one population of interest. Is almost like creating clusters. Then, with flags and a little cluster construction logic in place, the summative variables you create could be more valid, predictive and interpretable. The result, in my opinion, relative to imputing and replacing data points, is that you as the modeler aren't biasing the data based on your preconceived notions of what it should look like; or by the imputation procedure selected; or by what the modeler might consider big or small extreme scores; and not subject to arbitrary contextless cutoff values. Hope this makes a degree of sense.

Modeling Steps

Please complete the following steps for Task 3:

i. Go back to the original MONEYBALL data and start fresh with a clean slate. Transform the production variables (i.e. hits, strikeouts, etc) to production per game variables. Note that variables like ERA are already in this form.

ii. Do some internet searching and find evidence that allows you to put bounds on what modern era statistics are like per game in each of the production variables in the MONEYBALL dataset. This does not have to be perfect or exact. Good enough for government work is a nice rule of thumb. In other words, just be in the contextual "ball park" and not out of line.

iii. Create a flag variable for each production per game variable to identify records that are consistent with the range of values in modern era baseball for that variable. (0=consistent with modern era, 1=out of range of modern era)

iv. Create logic that looks across all the flag variables to identify those records that are consistent with modern era baseball across all production per game variables. Use this logic to create a categorical variable called MODERN, where 1 = consistent with modern era baseball, 0 = otherwise.

v. Retain your R code for creating the flags and the MODERN variable from steps iii) and iv) for exporting purposes in TASK 4. SAVE your R Code!

vi. Pull off only the records where MODERN=1. These are the records consistent with modern era baseball. Develop an OLS regression model to predict TARGET_WINS for all records where MODERN=1. For all other records (i.e. if MODERN=0) find the mean of Y. Use this mean as the predicted value for the model for these records. In other words, your model is in 2 parts:

    If MODERN=1, Y_contextual = b0 + b1*X2 + b2*X2 + ……..

    If MODERN=0, Y_contextual = mean of Y

    Call this model, the CONTEXTUAL Model.

vii. *Deliverables – Task 3:* Your project report should have a section that is headed as TASK 3 – DEVELOPING A CONTEXTUAL MODEL. This section of the project report should contain:

1. *Description of the procedure and details you used for obtaining the CONTEXTUAL model. This should include any criteria you used.*

2. *Report the CONTEXTUAL model, interpret the coefficients, and report the goodness of fit statistics.*

TASK 4: EXPORTING AND COMPARING MODEL PERFORMANCE (50 points)

You now have developed 3 models. Let's see how well these models perform in terms of exporting to a brand new set of data. This is also called Model Deployment and Generalization of the Model. In addition, I'll have you put together 2 additional models for comparative purposes. These won't be hard – trust me! In preparation, you will need the R code you developed in Tasks 1 and 3 that had anything to do with data cleaning, transformation, the creation of variables, flags, and conditions/logic.

Exporting, or deploying, a model is the process of using a previously developed model and hard coding the prediction equation (regression model) to predict Y for another set of data. This task will walk you through the process. If you are not quite comfortable with this idea yet, please see the Unit 1 Power Point file on Model DEPLOYMENT. We are going to deploy each model, then set up a way to compare the models and their predictive performance.

Please complete the following steps for Task 4:

i. Open the original MONEYBALL.CSV data file or the R version of it that you saved. You need to find 3 numbers: a) mean of TARGET_WINS(Y), b) the minimum value of TARGET_WINS(Y), and c) the maximum value of TARGET_WINS(Y). Write these numbers down so you have them easily available to you. Close the original MONEYBALL file. You don't need it anymore.

ii. Open the file MONEYBALL_EXPORT.CSV Notice, the sample size n and write that number down so you have access to it easily. Note, that this dataset contains all of the explanatory variables of the MONEYBALL dataset, but does NOT contain TARGET_WINS(Y). Read this file into RStudio.

iii. Create a new variable Y_mean = mean of TARGET_WINS(Y) which you obtained in step i-a) above. This is the predicted value of a mean only or the constant only model. Append Y_mean to the MONEYBALL_EXPORT dataset. The R-code for this model should look like: Y_mean <- 87.56

iv. Use n (the number of records in MONEYBALL_EXPORT), plus the minimum and maximum value of TARGET_WINS(Y), and the uniform random number generator in R to create a new variable called Y_RANDOM. Use the following R code:

y_random <- runif(n, minimum, maximum)

This will generate a column vector, Y_random, that is a random collection of numbers between the minimum and maximum values for Y. This isn't much of a model, but will be a good benchmark for performance purposes. Append Y_random to the MONEYBALL_EXPORT dataset.

v. This step is the hard part. What you did for model development has to be replicated for model deployment, in terms of the code you used to create and clean variables. You need to clean the data and construct the variables and flags, in exactly the same way you when you were developing the 3 models. These new variables and flags, plus the cleaning need to be implemented on the MONEYBALL_EXPORT data. We will do this one model at a time. Please note, and this is ***VERY IMPORTANT,*** you MUST have a predicted value for every record in the MONEYBALL_EXPORT file. If you have records without predictions, it will blow up the performance of your model.

a) For your BASE Regression model: In TASK 1 – Step iv), you were told to save your R-code that you developed to handle impute missing values, handle extreme scores, transforming variables and such. Run this exact same code on the MONEYBALL_EXPORT data. The idea is to prepare this data for modeling just like you did with the original data. Exporting means you have to replicate data cleaning in exactly the same way, otherwise the BASE model will not export well because you won't have clean data and all variables constructed. When that is done, write R-code to create a new variable called: Y_BASE. Your code will implement the BASE Regression model you reported in TASK 1 – Step vii). Your R-code will look like:

Y_BASE <- 1.985 + 5.256*X1 + 4.323*X2 – 10.652*X3 + 1.211*LOG(X1) + ….

Obviously, your coefficients and variables will be different. Append Y_BASE to the MONEYBALL_EXPORT dataset.

b) For your PARSIMONY model: In TASK 2, we just reduced the number of variables in the model. So, for this model, all we have to do is write code to implement the Y_parsimony variable from TASK 2 – step iv). Your R-code will look like:

Y_PARSIMONY <- 2.185 + 3.256*X1 + 5.323*X2

Obviously, your coefficients and variables will be different. Append Y_PARSIMONY to the MONEYBALL_EXPORT dataset.

c) ***For your CONTEXTUAL model***: In TASK 3 – steps iii and iv), you developed the flag variables to identify records that were consistent with modern era baseball. Just like for the BASE Model, run this exact same code on the MONEYBALL_EXPORT data. The idea is to prepare this data for modeling by creating the flags needed by the model. When that is done, write R-code to create a new variable called: Y_CONTEXTUAL. Your code will implement the CONTEXTUAL Regression model you reported in TASK 3 – Step v). Your R-code will look like:

If MODERN=1, Y_contextual <- b0 + b1*X2 + b2*X2 + ……..

If MODERN=0, Y_contextual <- mean of Y

Obviously, your coefficients and variables will be different. Append Y_CONTEXTUAL to the MONEYBALL_EXPORT dataset.

vi. So, at this point, the MONEYBALL_EXPORT file should contain the original variables, all the transformed variables, plus the predicted values for the 5 models. These predicted values for the 5 models are the variables: Y_MEAN, Y_RANDOM, Y_BASE, Y_PARSIMONY, and Y_CONTEXTUAL.

We no longer need all these variables from the MONEYBALL_EXPORT file, so let's pull off only those variables we need to complete the comparison. From the MONEYBALL_EXPORT file, extract the following variables and create a new dataframe called MODELCOMPARE. The variables needed are:

INDEX
Y_MEAN
Y_RANDOM
Y_BASE
Y_PARSIMONY
Y_CONTEXTUAL

vii. Almost done! The predicted values are nice and useful in practice. I hope you can explain why! But, for us to see how well the models performed, we need the actual value for Y we were trying to predict. So, import the file MONEYBALL_ACTUAL. This file contains only 2 variables: INDEX and Y_ACTUAL. If you look at this data, the INDEX values should be the same as in the dataframe you just constructed in step vi. INDEX should be sorted from smallest to largest for both files. If this is the case, and you have predicted values for each of the index values, then you may proceed with the comparison. If you are missing predicted values, then you have a problem and you'll have to go back and track down where it occurred and fix it. Then redo all the subsequent steps.

Merge the two files so that you have predicted and actuals in the same dataframe. You can do this in R or if you are having trouble figuring out how to do this, convert to EXCEL.CSV files and merge using EXCEL.

viii. We need to create a summative statistic for comparative purposes to evaluate the prediction performance of the 5 models. For each of the models, calculate the absolute deviation: ABS(Y_ACTUAL minus Y_predicted). Please note, you will substitute Y_MEAN, Y_RANDOM, etc. for Y_predicted. Furthermore, ABS is the function form for absolute value (i.e. positive deviation).

ix. Calculate the average absolute deviation using the deviations you created in step viii) for each of the 5 models.

x. Summarize the results of the 5 models in a table. Write a discussion these results in terms of model performance.

xi. *Deliverables – Task 4:* Your project report should have a section that is headed as TASK 4 – COMPARING MODEL PERFORMANCE. This section of the project report should contain:

1. A table summarizing the prediction results for the 5 models in terms of average absolute deviation. Please note that a perfect fitting model will have average absolute deviation of 0.
2. Discuss the model performance relative to the benchmarks of a perfect model, a random model, and a mean (constant term only) model. What can you infer about modeling from these results?

TASK 5: YOU ARE UP TO BAT (50 points)

Tasks 1 through 4 have been very prescriptive about what to do in guiding you through a small number of different modeling approaches. These Tasks were not intended to tell you what exactly to do, but rather to get you to think about modeling in different ways, as well as to introduce you to the ideas of model deployment or exporting. Hopefully, you gained some insight and perspective about modeling from those tasks.

There is not one perfectly correct way to approach model building. There are as many different perspectives as there are modelers out in the world. From your prior experiences in 410 or professionally, you may have an approach or perspective on modeling the MONEYBALL data that is very different than the three models given above. But, as all predictive modelers come to learn, when you are given a modeling task, it is your responsibility to complete it. So, in this spirit, you are now charged with the task of producing your best predictive model for the MONEYBALL data. This is an open ended task where you are free to do whatever you wish in modeling this data. You are charged with predicting TARGET_WINS using the variables in the MONEYBALL database.

Please complete the following steps for Task 5:

i. Report and discuss any additional EDA or data preparation you perform, over an above what has already been done in Tasks 1-4.

ii. Describe the modeling approach you took. 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, multivariate methods or whatever you think is best. But, ultimately, you need an OLS regression model as the prediction equation.

iii. 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.

iv. 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?

v. Write a Stand Alone R program that will export your model and predict the number of TARGET_WINS for the MONEYBALL_EXPORT.CSV data.

vi. Apply the R program to the data file MONEYBALL_EXPORT. Create a variable of predicted values for Y. Call this variable Y_yourlastname. Create a dataset that contains two variables: INDEX and Y_yourlastname. Please save this dataset as an EXCEL spreadsheet. Call this EXCEL spreadsheet Project1_yourlastname. ***Be sure you have a predicted value for every record!***

vii. *Deliverables – Task 5:* Your project report should have a section that is headed as TASK 5 – MY MODEL. This section of the project report should contain:

1. Written description of your Task 5 modeling approach and results.

2. The file: Project1_yourlastname for comparative grading. This file must be uploaded to CANVAS in the Project 1 DropBox.

WRAP UP OF PROJECT 1

- Your write up for Project 1 should be saved in PDF Format. Your write up should have five sections, one for each Task. You should simply take the Task Deliverables and

merge them all together into one document. Then please submit this project document to the Project 1 Dropbox by the due date.

- Please submit the PROJECT1_yourlastname EXCEL file to the Project 1 Dropbox.

CONGRATULATIONS! You are done with Project 1. Two more projects to go!