POINTS CLAIMED IN THIS ASSIGNMENT

Write Up	150
Stand along scoring program	50
Scored data file	50
Decision tree - bingo bonus	20
Pearson Corrleation Matrix Plots in R and Missingmap Graphical Analysis in R – bingo bonus	20
PROC GLM	10
Total	**300**

BASEBALL PREDICTIVE MODEL REPORT

INTRODUCTION

The purpose of this report is to present the analytical model created to predict the total wins per game season for a professional baseball team. The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology (figure 1) is followed during the model building process.

This report is organized according to the steps in the diagram, from *Business Understanding* to *Evaluation*. Exploratory Data Analysis (EDA) is performed first which leads to certain data preparation steps prior to linear regression model building. Multiple types of modeling techniques are investigated before a final model is selected according to certain predetermined performance criteria.

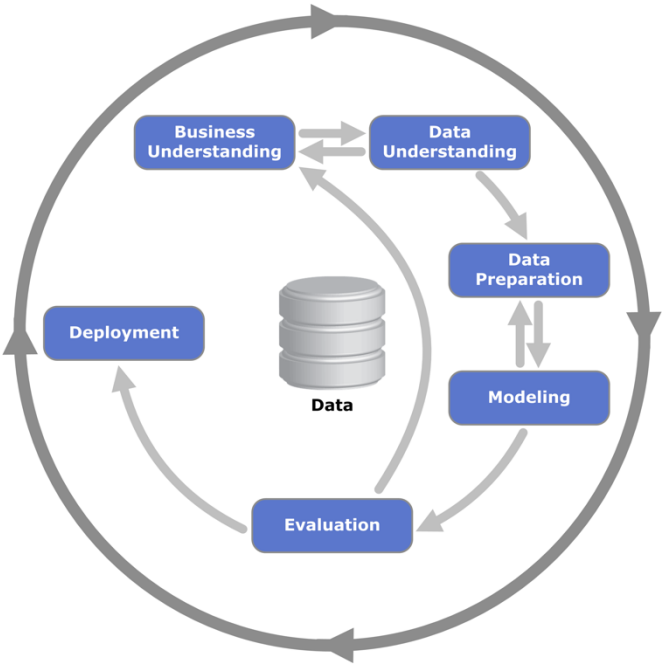


FIGURE 1. CRISP-DM METHODOLOGY

DATA UNDERSTANDING

Data from the years 1871 to 2006 is studied and used to build the predictive model. Each observation in the dataset is the performance of a team for a given year, with all of the statistics adjusted to match the performance of a 162 game season. The raw input data has 2276 observations, with the following variables available for analysis divided into three categories:

TABLE 1 INPUT DATA VARIABLES AND CATEGORIES

Category	Name	Description
BATTING	TEAM_BATTING_H	Base Hits by batters
	TEAM_BATTING_2B	Doubles by batters
	TEAM_BATTING_3B	Triples by batters
	TEAM_BATTING_HR	Homeruns by batters
	TEAM_BATTING_BB	Walks by batters
	TEAM_BATTING_SO	Strikeouts by batters
	TEAM_BASERUN_SB	Stolen bases
	TEAM_BASERUN_CS	Caught stealing
	TEAM_BATTING_HBP	Batters hit by pitch
PITCHING	TEAM_PITCHING_H	Hits allowed
	TEAM_PITCHING_HR	Homeruns allowed
	TEAM_PITCHING_BB	Walks allowed
	TEAM_PITCHING_SO	Strikeouts by pitchers
FIELDING	TEAM_FIELDING_E	Errors
	TEAM_FIELDING_DP	Double Plays

DATA INSIGHTS

Univariate data analysis of all the available data shows that the response variable (TARGET_WINS) the models will be trained to is well behaved. The data is normally distributed with a very slight left skew. There are no large outliers identified visually.

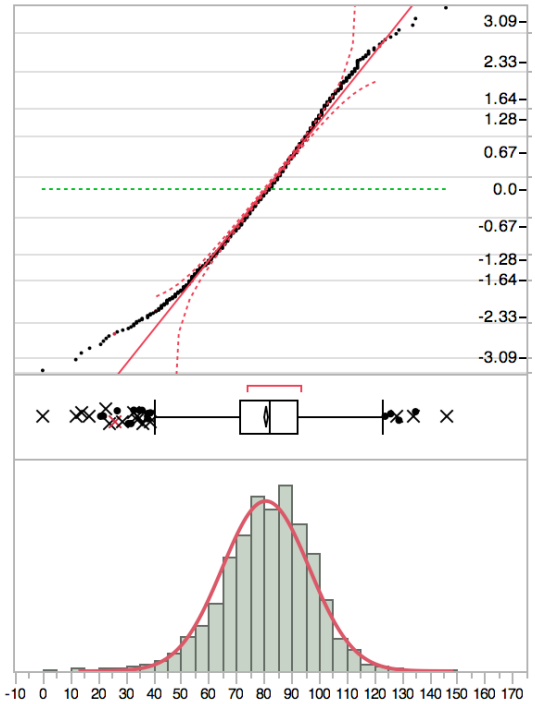


FIGURE 2 DISTRIBUTION AND NORMALITY PLOT FOR TARGET_WINS RESPONSE VARIABLE

Further univariate data analysis shows some variables have a bi-modal distribution, two of which are shown in figure 3. Bi-modal indicate the presence of an external force on the variable – perhaps relationship with some parameter present in the dataset, or even an unmeasured parameter. A direct relationship explaining the bimodality is not found in the dataset. So, transformations (log, and cube root) are investigated during the model building phases.

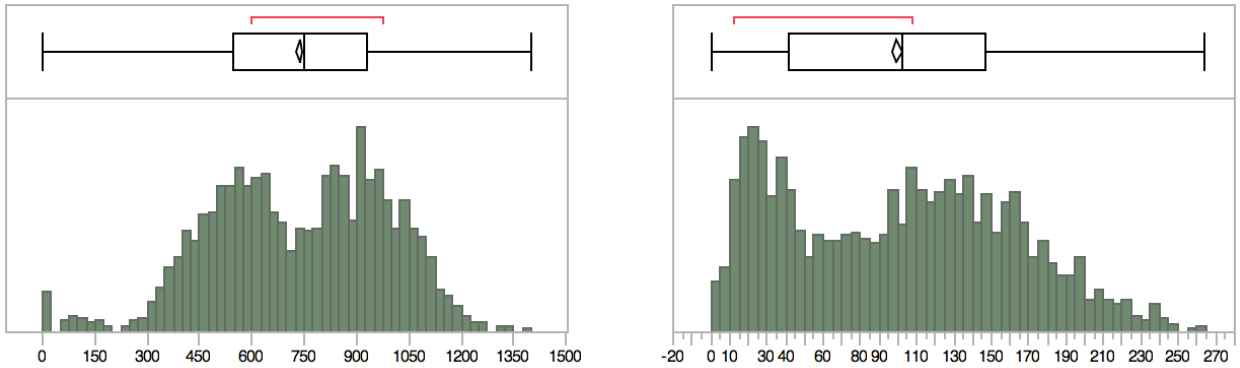


FIGURE 3 BATTING STRIKEOUTS (LEFT) AND BATTING HOME RUNS (RIGHT)

HIGHLY CORRELATED DATA

It is important to understand the correlation between the predictor and response variables. If there are variables which are singularly highly correlated to Target Wins, these variables can be good predictors of the wins. On the other hand, if there are variables which have a strong correlation between each other, there is an opportunity to avoid multicollinearity issues, reduce the total number of regressor variables by elimination or try derived variables.

Figure 4 shows the Pearson's correlation coefficients between all the variables. The variables have been intelligently grouped together using hierarchical clustering order to aid in visual identification of patterns. As we can see, the batting and pitching hit rates are highly correlated, as are the batting and pitching strike out rates. Stolen bases and Caught stealing are also highly correlated. Similarly, there is moderately high correlation between the batting and pitching hit rates, as well as the Walks variables.

During the model building process, the Variance Inflation Factors (VIF) for these variables were monitored to ensure the model did not suffer from multicollinearity issues.

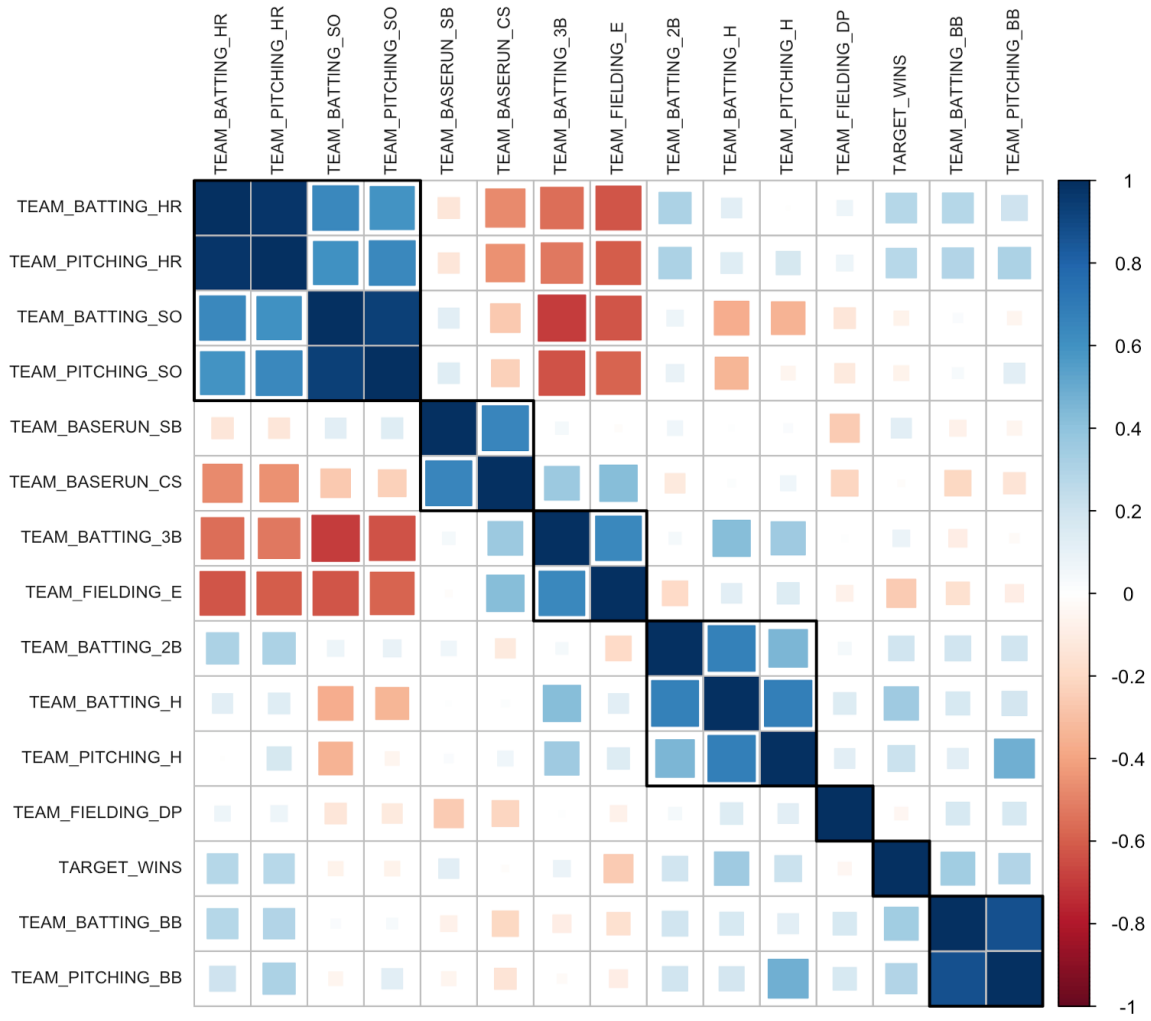


FIGURE 4 PEARSON'S CORRELATION COEFFICIENT MATRIX VISUALIZED

MISSING DATA

Six variables have missing data, as shown in the table below. Linear regression modeling cannot accept variables with missing data. All but one of the variables can be salvaged using imputation techniques. The variable BATTING_HBP has a missing rate of 92%. This is too large to derive any meaningful information from the data points available. The missing data for the remainder of the variables is taken care of by imputation techniques described in a later section.

TABLE 2 MISSING DATA PERCENTAGES

Variable	Label	% Missing	Salvageable?
TEAM_BATTING_HBP	Batters hit by pitch	92%	N
TEAM_BASERUN_CS	Caught stealing	34%	Y

TEAM_FIELDING_DP	Double Plays	13%	Y
TEAM_BASERUN_SB	Stolen bases	6%	Y
TEAM_BATTING_SO	Strikeouts by batters	4%	Y
TEAM_PITCHING_SO	Strikeouts by pitchers	4%	Y

Patterns of the missing data are investigated using graphical techniques. The plot in figure 5 shows the variables in decreasing order of missing % (missing values are coloredyellow). There is a discernable repeating pattern in the data highlighted by the blue boxes. Although temporal data is missing from the dataset, it is possible this repeating pattern represents the time period between 1871 to 2006. This information – leveraged by engineered variables – has proven valuable to development of a high quality model.

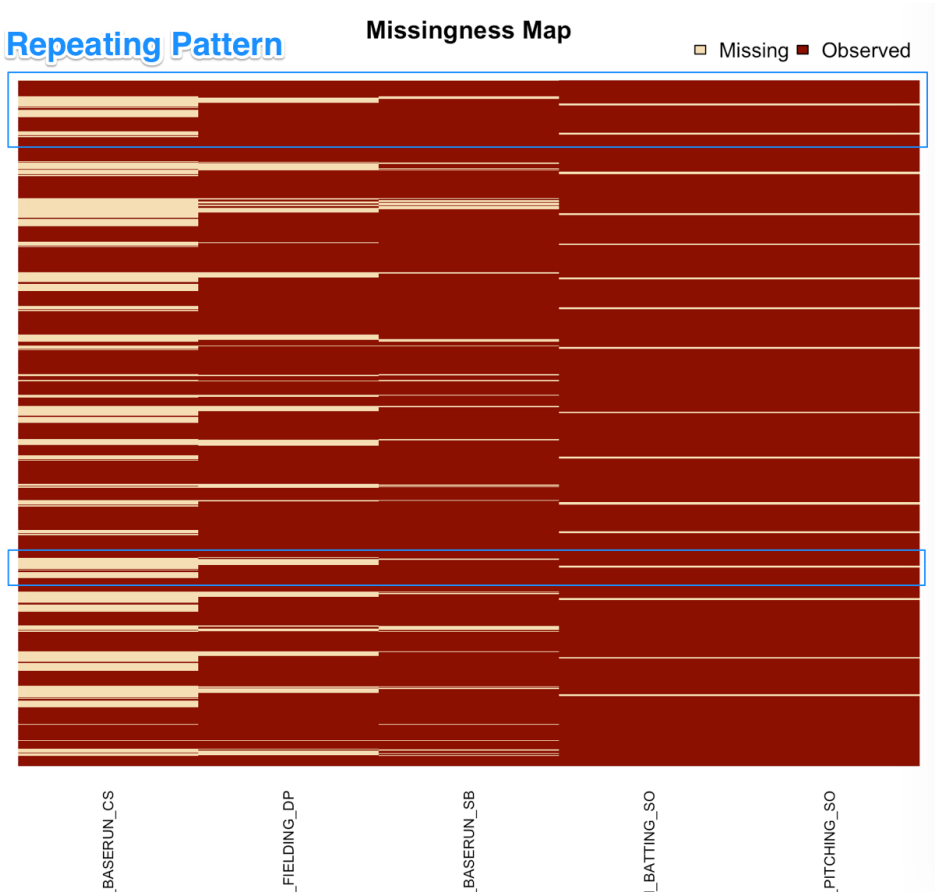


FIGURE 5 MISSINGNESS MAP FOR 5 VARIABLES

DATA PREPARATION

Based off of the EDA, three sets of data preparation steps are taken:

1. Data Imputation – Replace missing values with values computed by analyzing the remainder of the dataset
2. Outlier Elimination – Remove outliers based off of distribution of the data and performance of the regression models
3. Engineered Variables – Derive predictor variables based off of the original variables in the data set

DATA IMPUTATION

The method of data imputation selected for variables TEAM_BASERUN_CS, TEAM_FIELDING_DP, TEAM_BASERUN_SB, TEAM_BATTING_SO and TEAM_PITCHING_SO is the usage of decision trees. A decision tree is a tool that uses a tree-like graph of decisions and their possible consequences. This allows an analyst to more intelligently substitute missing data. An example of a decision tree for TEAM_PITCHING_SO is shown in figure 7.

As one can see from the 3-level tree, PITCHING_SO is much better imputed when the corresponding values of BATTING_SO and PITCHING_H are also considered. For each bucket at the tail end of the tree, the mean value of PITCHING_SO is substituted wherever there is missing data.

For each tree, the Split Best option was used to evaluate the splits, till the R-Sq value for the tree was sufficiently high *and* any additional splits only increased the R-Sq value by a small proportion. Furthermore, a k-fold cross validation (k=5) was performed to evaluate the performance of the imputation strategy.

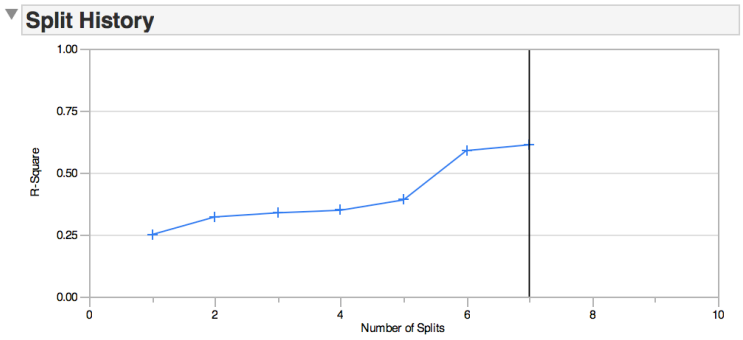


FIGURE 6 SPLIT-LEVEL SELECTION AND CROSSVALIDATION

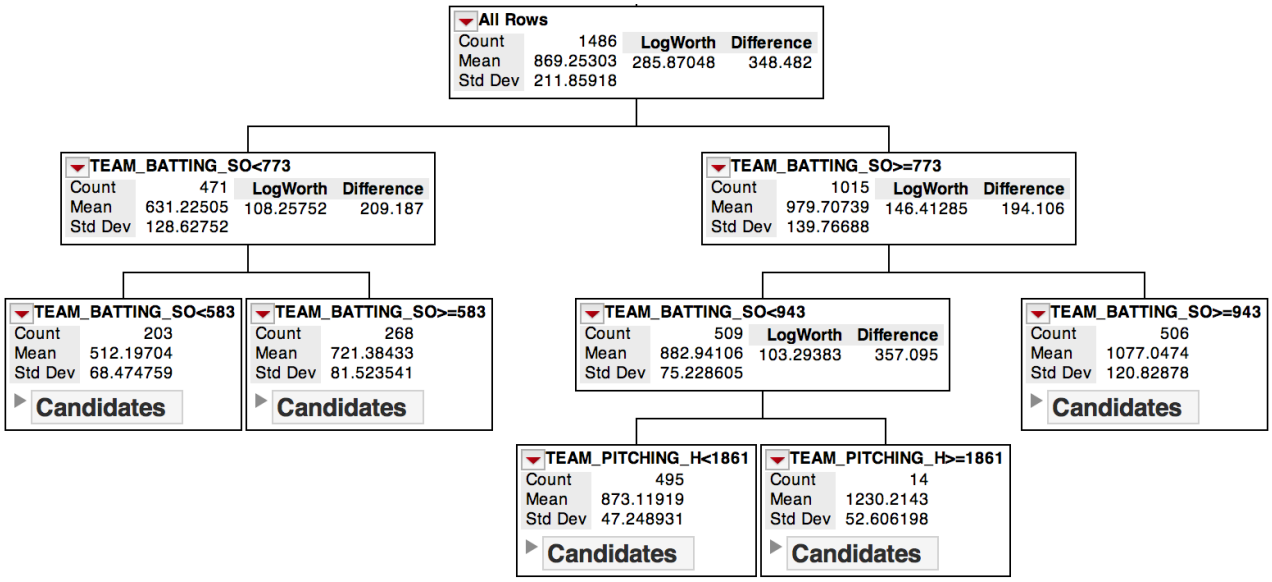


FIGURE 7 DECISION TREE FOR TEAM-PITCHING-STRIKE OUTS

Similar decision trees are performed on all variables with missing data.

OUTLIER ELIMINATION

Outliers adversely affect the fit of the regression model. Addressing outliers is an iterative process between the Data Preparation and Modeling steps. After many iterations of the modeling process, two methods are adopted to address outliers:

Step 1 – Eliminate data points that lie outside the 1% and 99% bounds for each variable. The cutoffs are shown in the table below.

Variable	1% Cutoff	99% Cutoff
TEAM_BATTING_H	-	1950
TEAM_BATTING_2B	141	352
TEAM_BATTING_3B	17	134
TEAM_BATTING_HR	4	235
TEAM_BATTING_BB	79	755
TEAM_BATTING_SO	200	1193
TEAM_BASERUN_SB	23	39
TEAM_BASERUN_CS	16	43
TEAM_FIELDING_E	86	1237
TEAM_PITCHING_B	-	924
TEAM_PITCHING_H	244	7093
TEAM_PITCHING_SO	-	1474

Step 2 – Eliminate influential outliers after regression modeling, as identified by the 'Leverage' calculations. The threshold for leverage points is selected to be 0.15.

These actions result in a reduction of data points from 2276 to 2223 or 2.4% reduction in number of observations.

ENGINEERED VARIABLES

The performance of the models is improved by creating new variables based off of the original dataset. These variables were evaluated in the modeling phase and retained if they added value to the final selected model.

TABLE 3 DERIVED VARIABLES

Engineered Variables	Formula	Interpretation
TEAM_BATTING_1B	TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR	How many single hits did the batters achieve?
SO_H_RATIO	$\frac{\text{TEAM\_BATTING\_SO}}{\text{TEAM\_BATTING\_H}}$	How many strike outs per total hits?
FIELDINGERRORS_PER_WALKSALLOWED	$\frac{\text{TEAM\_FIELDING\_E}}{\text{TEAM\_PITCHING\_BB}}$	How many fielding errors per walks allowed?
FIELDINGERRORS_PER_HITSGIVEN	$\frac{\text{TEAM\_FIELDING\_E}}{\text{TEAM\_PITCHING\_H}}$	How many field errors per hits allowed?
SB_PCT	$\frac{\text{TEAM\_BASERUN\_SB}}{\text{TEAM\_BASERUN\_SB} + \text{TEAM\_BASERUN\_CS}}$	What is the percentage of successful bases stolen of the total attempts?
PITCHING_SO_WALKS	$\frac{\text{TEAM\_PITCHING\_SO}}{\text{TEAM\_PITCHING\_BB}}$	How many strikeouts were taken during pitching per walks allowed?
TOTAL_TIME_ON_BASE	TEAM_BATTING_H + TEAM_BATTING_BB	What is the total time spent on the base?

Furthermore, five variables are created which indicate if the corresponding columns have missing data which need to be imputed: MISS_TEAM_PITCHING_SO, MISS_TEAM_BATTING_SO, MISS_TEAM_BASERUN_SB, MISS_TEAM_BASERUN_CS and MISS_TEAM_FIELDING_DP.

MODELING

MODEL BUILDING

Multiple versions of a multilinear regression model are built and examined at this step. First, models with only the original variables (post-cleansing) are used in the model. Thereafter, engineered variables and transformations are added to the model to evaluate the change in performance.

A summary of all the models evaluated is shown in the table below.

Concept#	Details	Outliers Eliminated	Selection Criteria	No of Variables	Adj Rsq	AIC	SBC	25% Test Data Error Score	VIFs > 10
6	Original & Engineered Variables	{1%-99%} & Stricter Leverage Point Removal	Model 5 + VIFs Addressed	17	0.4071	10737.61	10834.53	11.5327	7
5	Original & Engineered Variables	{1%-99%} & Leverage Points	Adj Rsq + Lowest SBC	17	0.4162	10791.51	10894.21	12.05017	5
4	Original & Engineered Variables	{1%-99%} & Leverage Points	Stepwise	17	0.4175	10703.72	11110	12.94765	6
1	Original Variables Only	{1%-99%} & Leverage Points	None	14	0.32075	11034.48	11120	13.89305	7
2	Original Variables Only	{1%-99%} & Leverage Points	Stepwise	14	0.32075	11034.48	11120	13.89305	7
3	Original Variables Only	{1%-99%} & Leverage Points	Adj Rsq + Lowest SBC	12	0.32064	11042.89	11111.31	15.72624	7

A number of techniques for variable selection were chosen to evaluate alternative models, mainly: Stepwise regression and Adjusted Rsquare sorted by lowest SBC value, since the preliminary data analyses showed these techniques to have the most promising results.

For each model, the AIC, SBC, Adjusted Rsquare, Number of Variables, and VIF performance is documented. Also, the performance of the model to a 25% test data, as measured by an error score is documented. This score gives an indication of the performance of the model on new-test dataset which the model has not seen before.

All the models evaluated pass the minimum requirements for a regression model: normal residuals, homoscedastic residuals, no patterns in the residual plots etc. These basic model adequacy checks were performed on all models.

All models evaluated had a minimum of 5 variables with Variance Inflation Factors (VIF) greater than 10. This typically indicates an issue of multicollinearity with the model. Though, when these variables were removed from the model, the performance of the model decreased significantly. Thus, it is decided to keep this variables in the model at this stage.

EVALUATION

SELECTION CRITERIA

The top 2 models as sorted by AIC, SBC and 25% test error criteria have 17 variables in the model. Thus, both models are equally parsimonious.

Concept #6 was chosen as the top performing model based on the lowest values of all performance indicators – AIC, SBC and 25% test error despite having a slightly lower Adjusted Rsquare value. This concept has the strictest criteria for outlier detection and elimination, which proved to be a significant factor in improving the model performance.

Concept #6 is selected as the final model and is explored below.

FINAL MODEL

The final linear regression model selected to predict the team victories is as follows:

Predicted Target	64.7454	+		
Wins =	0.0223	x	Total Time on Base	+
	0.1476	x	Triples by batters	+
	0.0959	x	Homeruns by batters	+
	0.0585	x	Walks by batters	+
	0.0915	x	Stolen Bases	+
	0.0115	x	Strikeouts by Pitchers	+
	5.9201	x	Missing Data – Strikeouts by Pitchers	+
	53.9604	x	Missing Data – Stolen Bases	+
	4.0409	x	Missing Data – Caught Stealing	+
	7.8913	x	Missing Data – Double Play	+
	- 0.0562	x	Fielding Errors	+
	- 0.0906	x	Double Plays	+
	- 0.0303	x	Strikeouts by batters	+
	- 0.0380	x	Doubles by batters	+
	- 0.043	x	Walks Allowed	+
	- 94.4698	x	Fielding Errors Per Hits Given	+
	- 11.5830	x	Stolen Base Percentage	

The performance of the model is the highest when the predicted Target Wins is capped between 20 and 115.

This equation does make very intuitive sense for the variables selected in the model. The first few variables in the equation – all positive – Total Time on base, Triples by Batters, Homeruns by Batters, Walks by Batters and Stolen Bases – all indicate the performance by the team's offensive. The better the batters do, the higher the chances of winning the game is. Similarly, the larger the strikeouts by pitchers, the higher the chances of winning games.

On the other hand, the variables – Fielding Errors, Strikeouts By Batters, Walks Allowed, and Fielding Errors Per Hits Given – all negatively affect the changes of winning a game. These are appropriately negatively weighted.

Three variables which have counter-intuitive weighting are Double Plays, Doubles by Batters and Stolen Base Percentage. Intuitively, all three should have positive weights. Decrease in the changes of wins due to increase in Doubles by batters and Stolen Base Percentages can perhaps be attributed to an increase in the risks being taken on the field by the batters. The increase in aggressive behavior by batters perhaps results in higher mistakes overall which reduces the chances of wins. Similarly, on Double Plays, it's possible that as the defense gets more aggressive, it offers greater opportunities for mistakes and reduces the chances of wins.

Finally, the interesting variables in the equation are the "Missing" flags, which are identified as important to maintain the accuracy of the model – especially Missing Data – Stolen Bases, which if data is present, increases the wins by 54 games. The hypothesis for why these flags are important is there has been some temporal changes in the game from 1871 to 2006. These could include changes in the way the game is played, the way statistics are calculated or saved etc. The author has decided to retain these variables given their positive contribution, but also recommends consulting with subject matter experts to determine the causality between the variables and the predicted wins.

CONCLUSION

Many models were successfully evaluated and a top performing model was selected using the Adjusted-Rsquare+SBC criteria combined with the performance on a 25% test dataset. The model performs well with a low error score on the test dataset, and it does not violate any assumptions required for linear regression modeling.

There are some variables in the final selected model which would benefit from further study with subject matter experts. This should be the next step of the process. Furthermore, the model should be monitored over time and the variables with high VIF values should be

investigated. It is possible the model may not perform well with completely new data given the higher confidence intervals of the regression coefficients. Regularly re-running the model will ensure it is behaving in a manner acceptable to the business needs.

BINGO BONUS

20 POINTS – PROC GLM & PROC GENMOD

PROC GLM

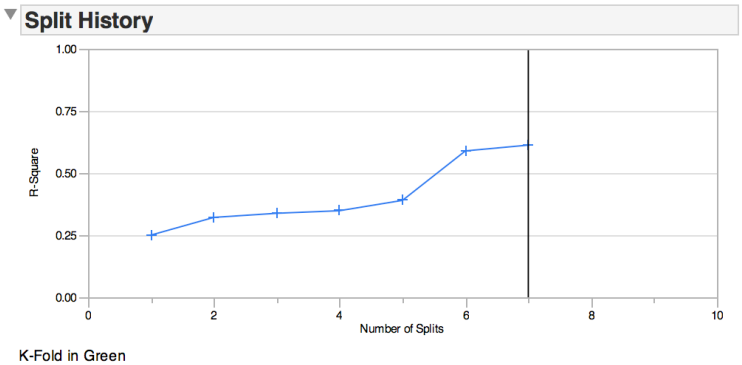
```
PROC GLM data=cleanfile;
    model TARGET_WINS=
    TEAM_BATTING_2B
    TEAM_BATTING_3B
    TEAM_BATTING_HR
    TEAM_BATTING_BB
    TEAM_BATTING_SO
    TEAM_BASERUN_SB
    TEAM_PITCHING_BB
    TEAM_PITCHING_SO
    TEAM_FIELDING_E
    TEAM_FIELDING_DP
    MISS_TEAM_PITCHING_SO
    MISS_TEAM_BASERUN_SB
    MISS_TEAM_BASERUN_CS
    MISS_TEAM_FIELDING_DP
    FIELDINGERRORS_PER_HITSGIVEN
    SB_PCT
    TOTAL_TIME_ON_BASE;
OUTPUT OUT=OUT;
run;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	53.54064968	7.03107371	7.61	<.0001
TEAM_BATTING_2B	-0.02846853	0.00866063	-3.29	0.0010
TEAM_BATTING_3B	0.13465628	0.01689128	7.97	<.0001
TEAM_BATTING_HR	0.08542913	0.00916153	9.32	<.0001
TEAM_BATTING_BB	0.05564175	0.01380321	4.03	<.0001
TEAM_BATTING_SO	-0.02827554	0.00714605	-3.96	<.0001
TEAM_BASERUN_SB	0.08605145	0.00661348	13.01	<.0001
TEAM_PITCHING_BB	-0.04339873	0.01092209	-3.97	<.0001
TEAM_PITCHING_SO	0.01366064	0.00643414	2.12	0.0339
TEAM_FIELDING_E	-0.05647447	0.00832620	-6.78	<.0001
TEAM_FIELDING_DP	-0.08465715	0.01294973	-6.54	<.0001
MISS_TEAM_PITCHING_S	6.00819826	1.46701933	4.10	<.0001
MISS_TEAM_BASERUN_SB	52.42985115	3.06235548	17.12	<.0001
MISS_TEAM_BASERUN_CS	3.95257025	0.87022743	4.54	<.0001
MISS_TEAM_FIELDING_D	7.42970202	1.63551550	4.54	<.0001
FIELDINGERRORS_PER_H	-72.24253255	16.39851156	-4.41	<.0001
SB_PCT	-11.09290937	4.72795449	-2.35	0.0191
TOTAL_TIME_ON_BASE	0.02541939	0.00445285	5.71	<.0001

20 POINTS – DECISION TREES USING SAS JMP

For all the data columns with missing data (TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_PITCHING_SO, TEAM_FIELDING_DP) the missing data was imputed using decision trees in SAS JMP.

For each tree, the **Split Best** option was used to evaluate the splits, till the R-Sq value for the tree was sufficiently high *and* any additional splits only increased the R-Sq value by a small proportion. Furthermore, a **k-fold cross validation** (k=5) was performed to evaluate the performance of the imputation strategy.

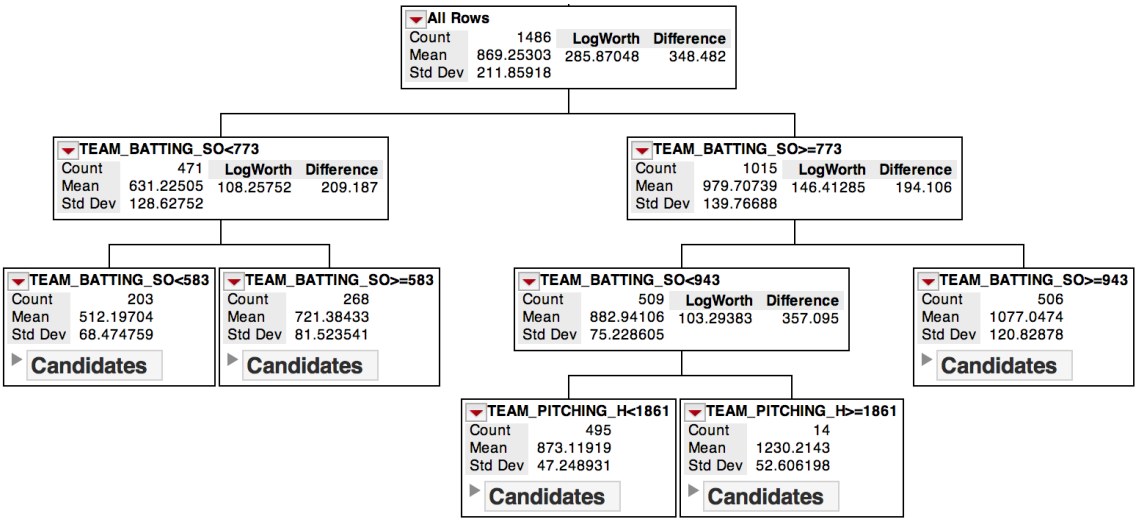


K-Fold in Green

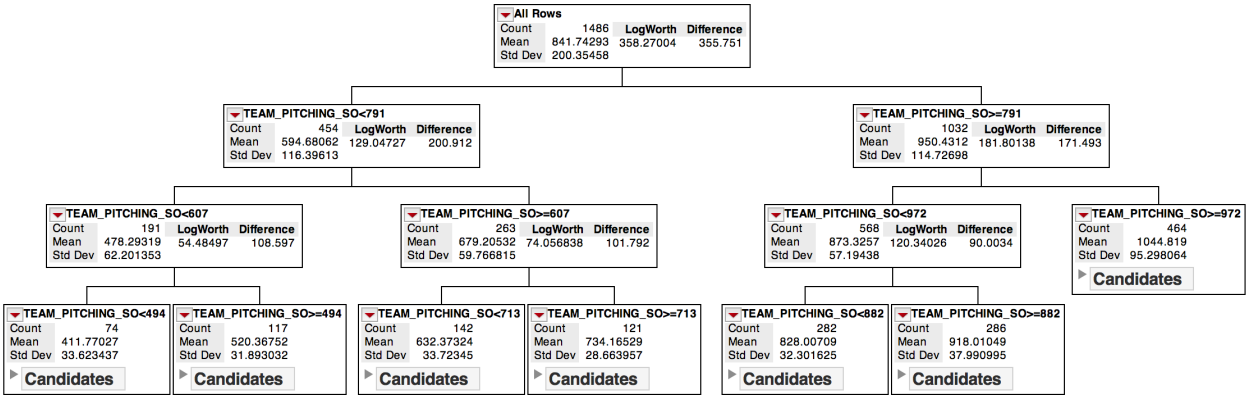
Crossvalidation

k-fold		SSE	RSquare
5	Folded	308076.12	0.6027
	Overall	300517.072	0.6124

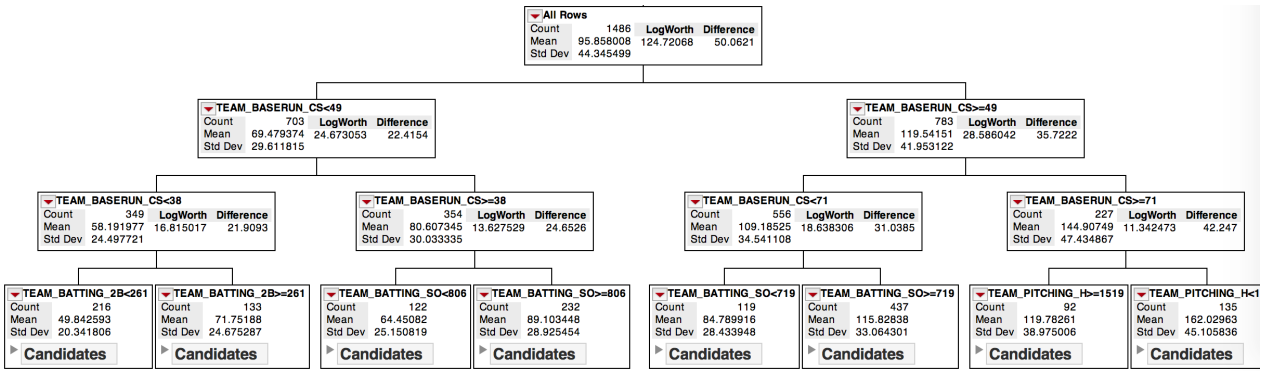
TEAM_PITCHING_SO



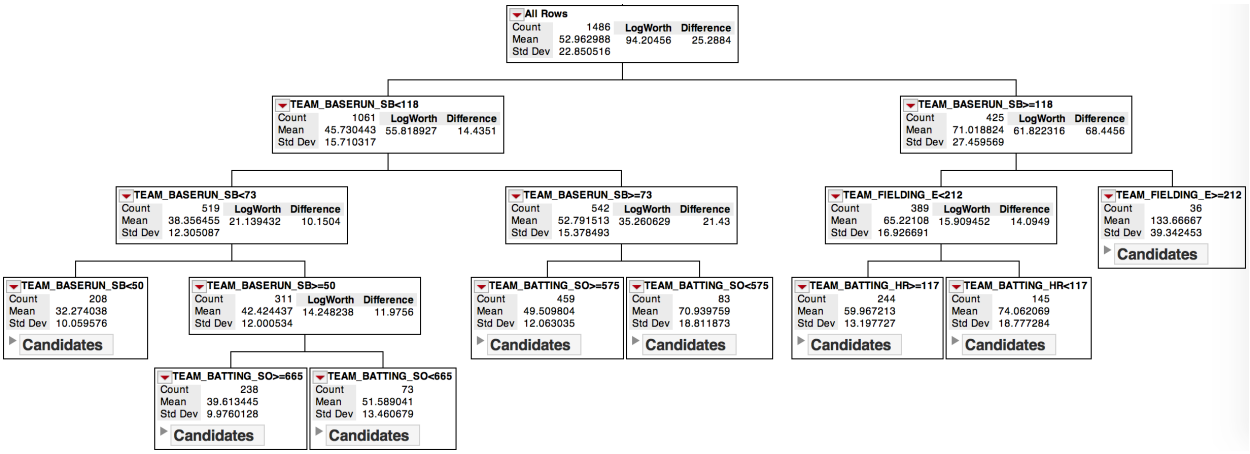
TEAM_BATTING_SO



TEAM_BASERUN_SB



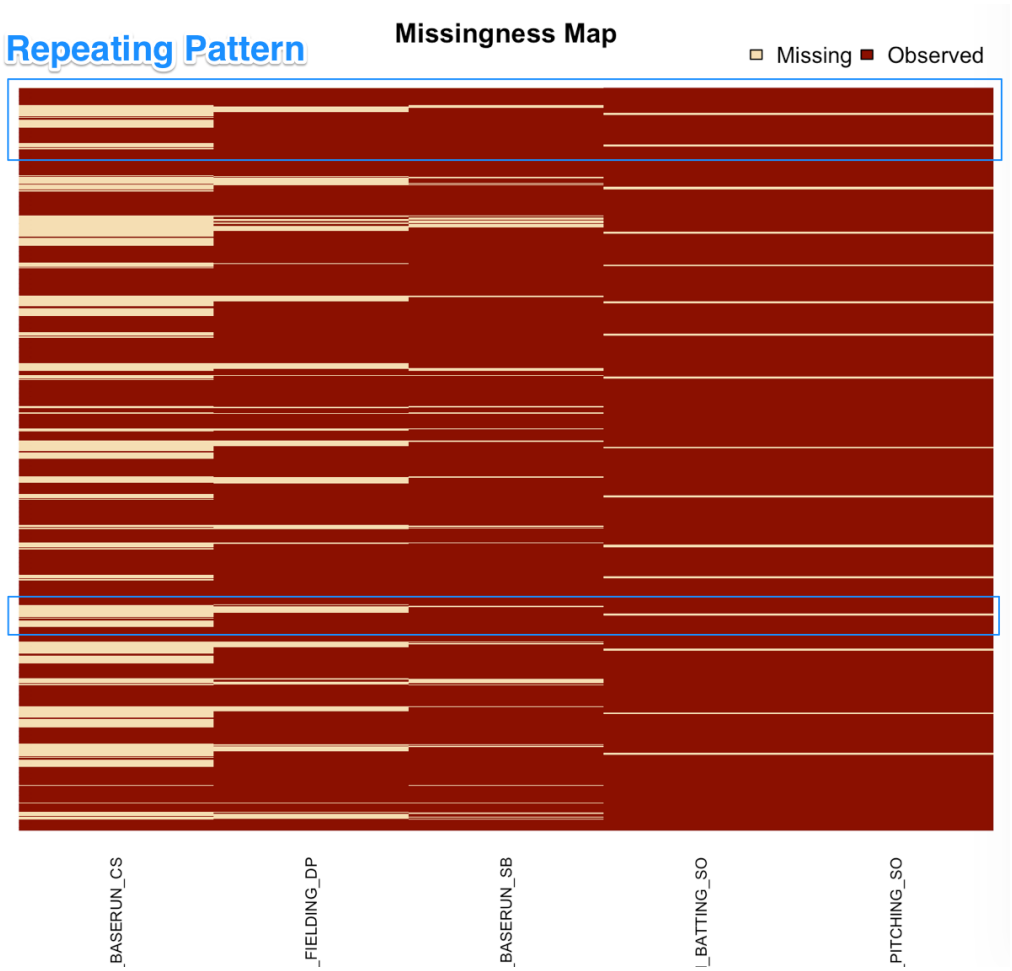
TEAM_BASERUN_CS



Once the decision trees were selected, they were transferred to SAS code which processed the input data as well as the test data.

EXTRA -- 20 POINTS – MISSING VALUE PATTERN IDENTIFICATION IN R

Using the Amelia package in R, we can evaluate patterns of the missing data. The output of Amelia is a **Missingness Map** shown below. It plots the variables in decreasing order of missing %; the missing values are color coded yellow. There is a clear repeating pattern in the data highlighted by the blue boxes. This information has been valuable and has been used in the main paper.


Repeating Pattern
Missingness Map
Missing
Observed
BASERUN_CS
FIELDING_DP
BASERUN_SB
BATTING_SO
PITCHING_SO