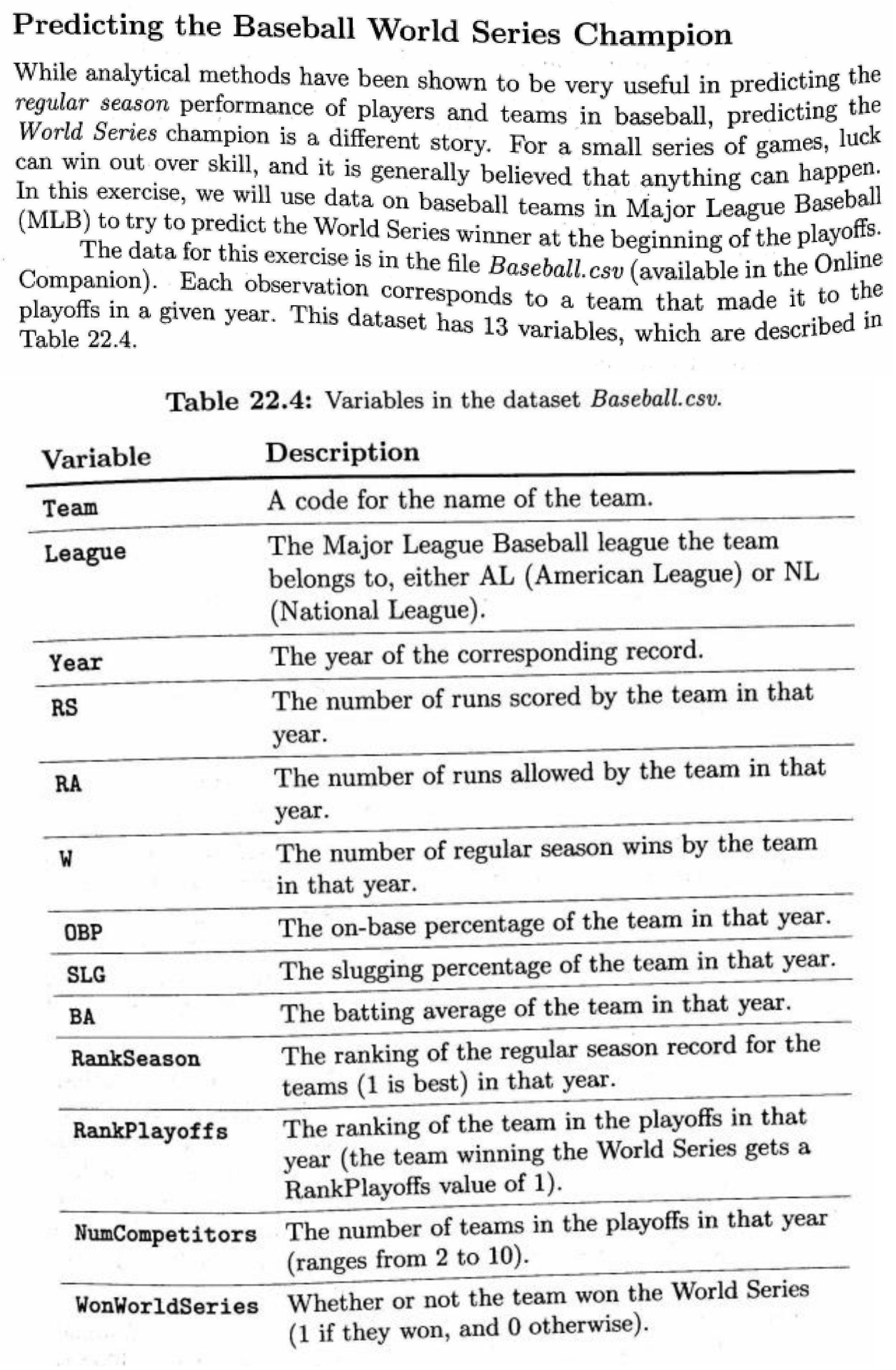
MSCI 433 (W2022) Assignment 2

Due Date: Feb 11, 11:30pm (EDT)

Problem 1.



Predicting the Baseball World Series Champion

While analytical methods have been shown to be very useful in predicting the *regular season* performance of players and teams in baseball, predicting the *World Series* champion is a different story. For a small series of games, luck can win out over skill, and it is generally believed that anything can happen. In this exercise, we will use data on baseball teams in Major League Baseball (MLB) to try to predict the World Series winner at the beginning of the playoffs.

The data for this exercise is in the file *Baseball.csv* (available in the Online Companion). Each observation corresponds to a team that made it to the playoffs in a given year. This dataset has 13 variables, which are described in Table 22.4.

Table 22.4: Variables in the dataset *Baseball.csv*.

Variable	Description
Team	A code for the name of the team.
League	The Major League Baseball league the team belongs to, either AL (American League) or NL (National League).
Year	The year of the corresponding record.
RS	The number of runs scored by the team in that year.
RA	The number of runs allowed by the team in that year.
W	The number of regular season wins by the team in that year.
OBP	The on-base percentage of the team in that year.
SLG	The slugging percentage of the team in that year.
BA	The batting average of the team in that year.
RankSeason	The ranking of the regular season record for the teams (1 is best) in that year.
RankPlayoffs	The ranking of the team in the playoffs in that year (the team winning the World Series gets a RankPlayoffs value of 1).
NumCompetitors	The number of teams in the playoffs in that year (ranges from 2 to 10).
WonWorldSeries	Whether or not the team won the World Series (1 if they won, and 0 otherwise).

a) Let us start by exploring our dataset.

i) Each row of *Baseball.csv* represents a playoff team's performance in a particular year. Through the years, different numbers of teams have been invited to the playoffs. How has the number of teams making it to the playoffs each year changed, according to this dataset?

ii) Given that a team has made it to the playoffs, it is much harder to win the World Series if there are 10 teams competing for the championship versus just two. Therefore, we have the variable **`NumCompetitors`** in our dataset. **`NumCompetitors`** contains the number of total teams making the playoffs in the year of the observation. For instance, **`NumCompetitors`** is 2 for the 1962 New York Yankees, but it is 8 for the 1998 Boston Red Sox. Without knowing anything else about the teams in the playoffs, can you think of a simple model that uses **`NumCompetitors`** to predict the probability of a team winning?

b) Let us now build logistic regression models to predict the World Series winner.

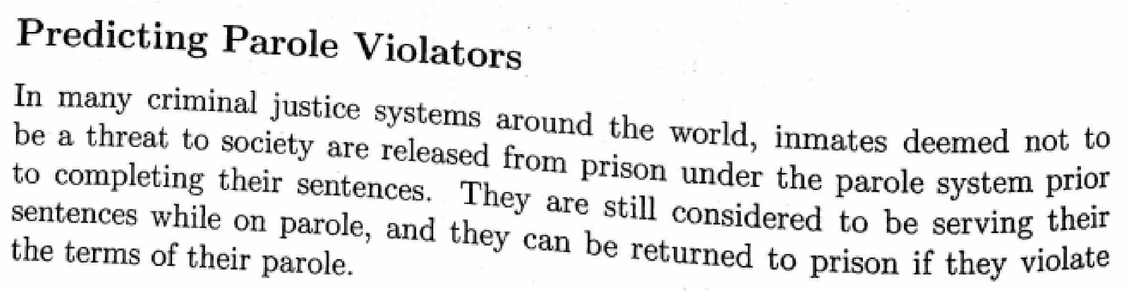
i) When we are not sure which of our variables are useful in predicting a particular outcome, it is often helpful to build bivariate models, which are models that predict the outcome using a single independent variable. Build a bivariate logistic regression model using each of the following variables as the independent variable to predict **WonWorldSeries**, and the entire dataset as the training set each time: **Year**, **RS**, **RA**, **W**, **OBP**, **SLG**, **BA**, **RankSeason**, **NumCompetitors**, and **League**. You should have created 10 logistic regression models. Describe each of the models by giving the regression equation and the accuracy of the model. For which models is the independent variable significant? In your opinion, which are the best models and why?

ii) Now, build a logistic regression model using all of the variables that you found to be significant in the bivariate models as the independent variables, and the entire dataset to train the model. Are all of the independent variables significant in this model? Why would some independent variables be significant in the bivariate model using that variable, but then not significant in a model that uses more than one independent variable? Be sure to provide numerical evidence for your claim.

iii) Using any number of the independent variables that you found to be significant in the bivariate models, find what you think is the best model, and justify why you think it is the best. How many independent variables are used in your final model?

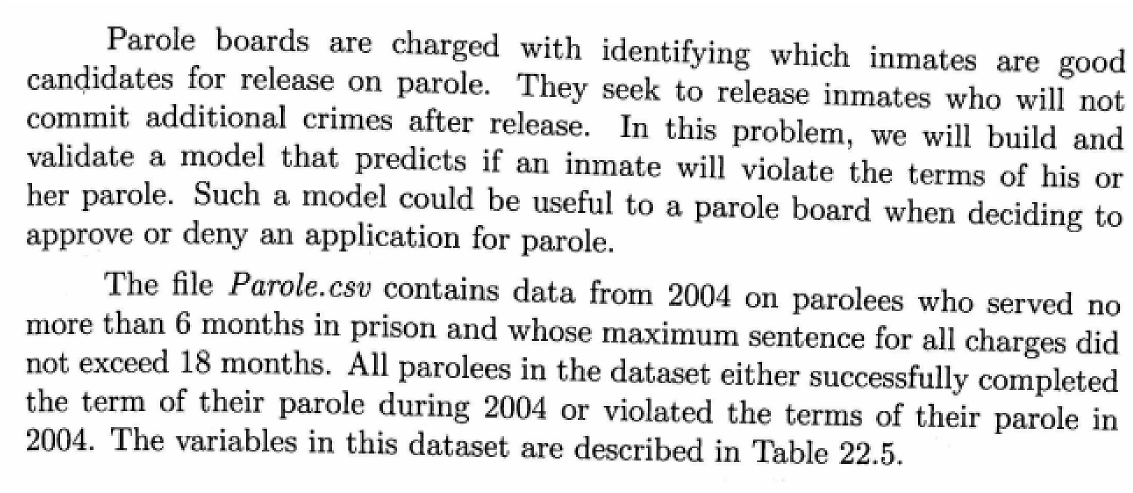
iv) Do your findings in this problem confirm or reject the claim that the playoffs is more about luck than skill? Why?

Problem 2.



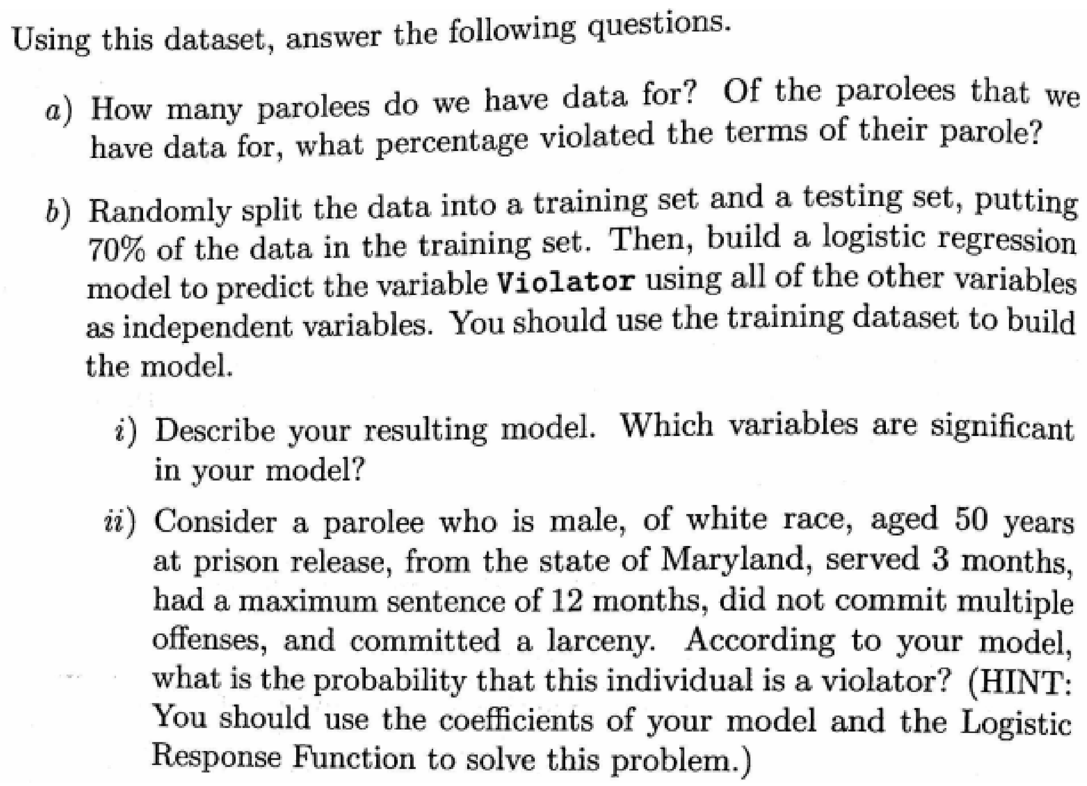
Predicting Parole Violators

In many criminal justice systems around the world, inmates deemed not to be a threat to society are released from prison under the parole system prior to completing their sentences. They are still considered to be serving their sentences while on parole, and they can be returned to prison if they violate the terms of their parole.



Parole boards are charged with identifying which inmates are good candidates for release on parole. They seek to release inmates who will not commit additional crimes after release. In this problem, we will build and validate a model that predicts if an inmate will violate the terms of his or her parole. Such a model could be useful to a parole board when deciding to approve or deny an application for parole.

The file *Parole.csv* contains data from 2004 on parolees who served no more than 6 months in prison and whose maximum sentence for all charges did not exceed 18 months. All parolees in the dataset either successfully completed the term of their parole during 2004 or violated the terms of their parole in 2004. The variables in this dataset are described in Table 22.5.



Using this dataset, answer the following questions.

a) How many parolees do we have data for? Of the parolees that we have data for, what percentage violated the terms of their parole?

b) Randomly split the data into a training set and a testing set, putting 70% of the data in the training set. Then, build a logistic regression model to predict the variable `Violator` using all of the other variables as independent variables. You should use the training dataset to build the model.

i) Describe your resulting model. Which variables are significant in your model?

ii) Consider a parolee who is male, of white race, aged 50 years at prison release, from the state of Maryland, served 3 months, had a maximum sentence of 12 months, did not commit multiple offenses, and committed a larceny. According to your model, what is the probability that this individual is a violator? (HINT: You should use the coefficients of your model and the Logistic Response Function to solve this problem.)

Table 22.5: Variables in the dataset *Parole.csv.*

Variable	Description
Male	1 if the parolee is male, and 0 if female.
RaceWhite	1 if the parolee is white, and 0 otherwise.
Age	The parolee's age in years when he or she was released from prison.
State	The parolee's state, either Kentucky, Louisiana, Virginia, or Other. These three states were selected due to having a high representation in the dataset.
TimeServed	The number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
MaxSentence	The maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
MultipleOffenses	1 if the parolee was incarcerated for multiple offenses, and 0 otherwise.
Crime	The parolee's main crime leading to incarceration, either Larceny, Drugs (representing a drug-related crime), Driving (representing a driving-related crime), or Other.
Violator	1 if the parolee violated their parole, and 0 if the parolee completed their parole without violation.

iii) Now compute the model's predicted probabilities for parolees in the testing set. Then create a confusion matrix for the test set using a threshold of 0.5. What is the model's false positive rate on the test set? False negative rate? Overall accuracy?

iv) Compare your accuracy on the test set to a *baseline* model that predicts every parolee in the test set is a non-violator, regardless of the values of the independent variables. Does your model improve over this simple model?

v) Consider a parole board that might use your model to predict whether parolees will be violators or not. The job of a parole board is to make sure that a prisoner is ready to be released into free society, and therefore parole boards tend to be particularly concerned about releasing prisoners who will violate their parole. Would the parole board be more concerned by false positive errors or false negative errors? How should they adjust their threshold to reflect their error preferences?

vi) Compute the AUC of the model on the test set, and interpret what the number means in this context. Considering the AUC, the accuracy compared to the baseline model, and what happens when the threshold is adjusted, do you think this model is of value to a parole board? Why or why not?

c) Our goal in this problem has been to predict the outcome of a parole decision, and we used a publicly available dataset of parole releases for predictions. It is always important to evaluate a dataset for possible sources of bias, especially when the dataset only contains a subset of the observations of interest.

The *Parole.csv* dataset contains all individuals released from parole in 2004, either due to completing their parole term or violating the terms of their parole. However, it does not contain parolees who neither violated their parole nor completed their term in 2004, causing non-violators to be underrepresented. This is called *selection bias* or *selecting on the dependent variable*, because we used our dependent variable (parole violation) to select only a subset of all relevant parolees to include in our analysis. How could we improve our dataset to best address selection bias?