NBA Final Project

by: Andrew Acosta

Introduction

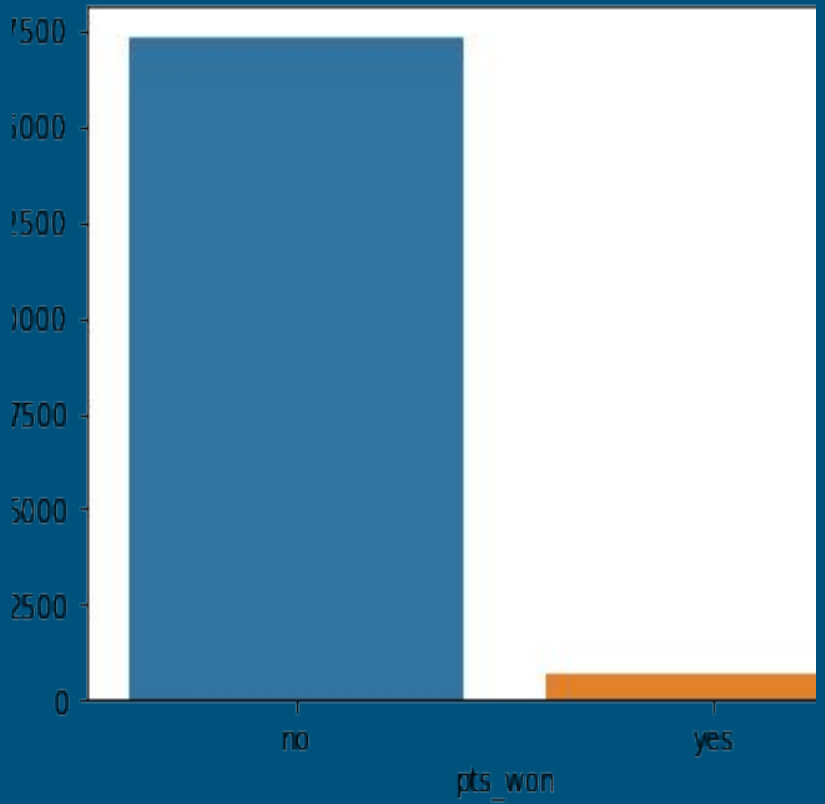
As a data analyst, I have chosen to do a classification project on an NBA player statistics dataset. This dataset holds all NBA player stats since 1980. One of the columns in this dataset contain information on whether or not a player received MVP points that season. The goal of this project is to predict what statistics of an NBA player affect whether or not they were voted at all for the MVP of the league.

Data exploration/cleaning

It all starts with understanding the data. In this first stage I must explore the dataset I’m working on and understand what each column brings to the table. I also must start cleaning the data so that I can actually work on it.

Visualizations

A part of understanding the data I'm working on is visualizing that data. I use different plots to visualize the comparison on my target column and the other columns. Here is one example of a visualization I used to visualize the value count difference in the target column "pts_won".



Target and feature selection

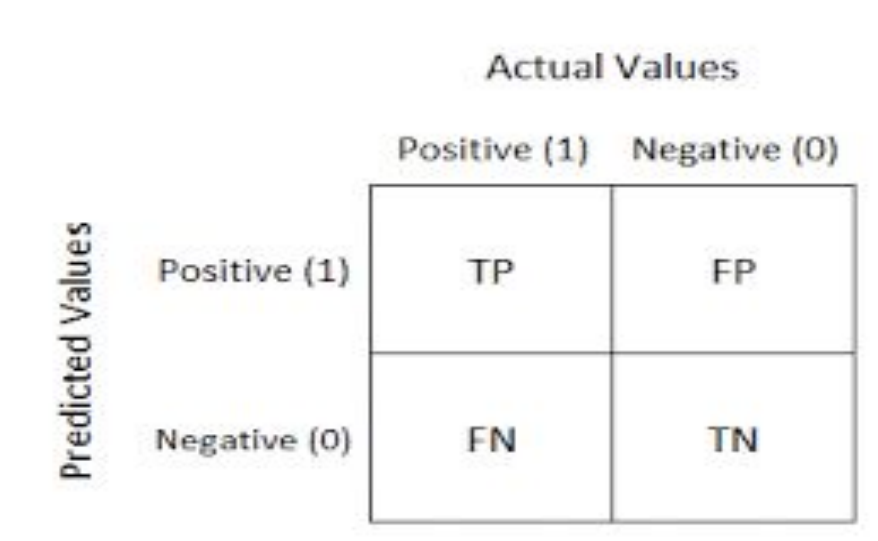
At this point I split the data into x and y columns. Then, from the x column I separate the numericals from categoricals to scale and encode. When done scaling and encoding, the results are concatenated together which makes up the feature columns. I now have my features ready. Then the target album needs to be balanced.

Performing models

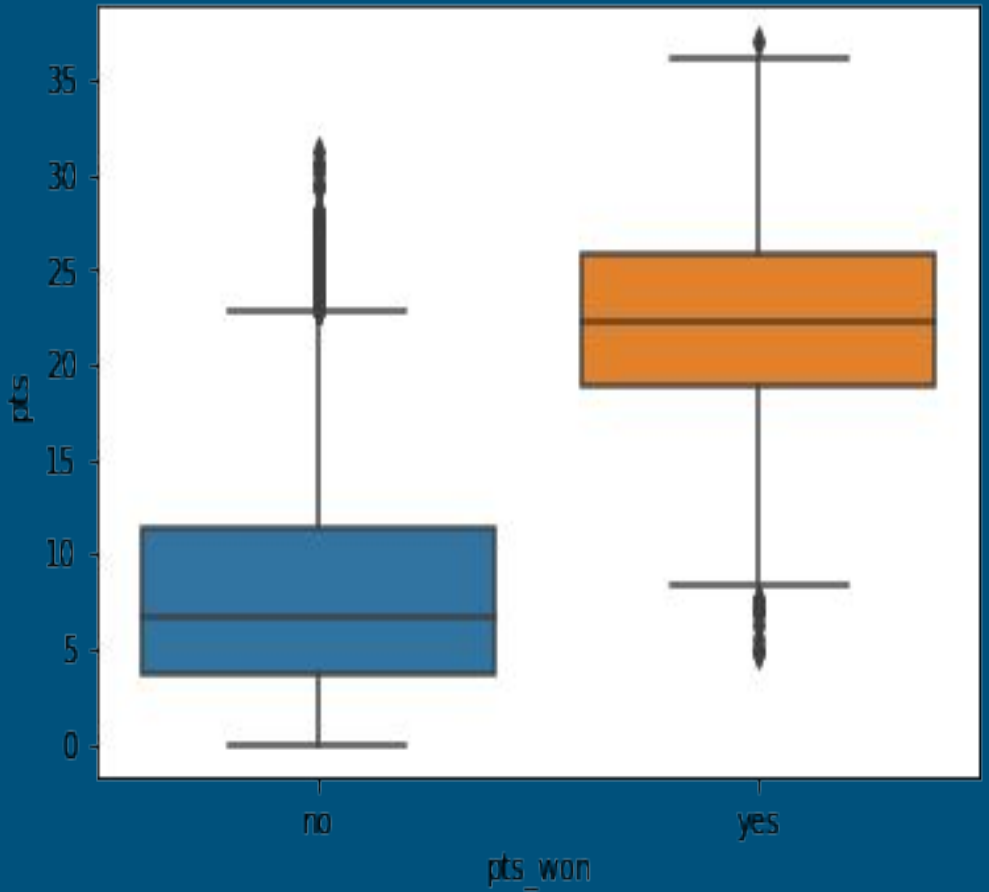
In order to predict which stats of a player directly affects whether or not they were voted for mvp, I must perform machine learning models and test their accuracy with the data I’ve prepared. The models I implemented were a logistic regression model, a decision tree model, and a random forest classifier model.

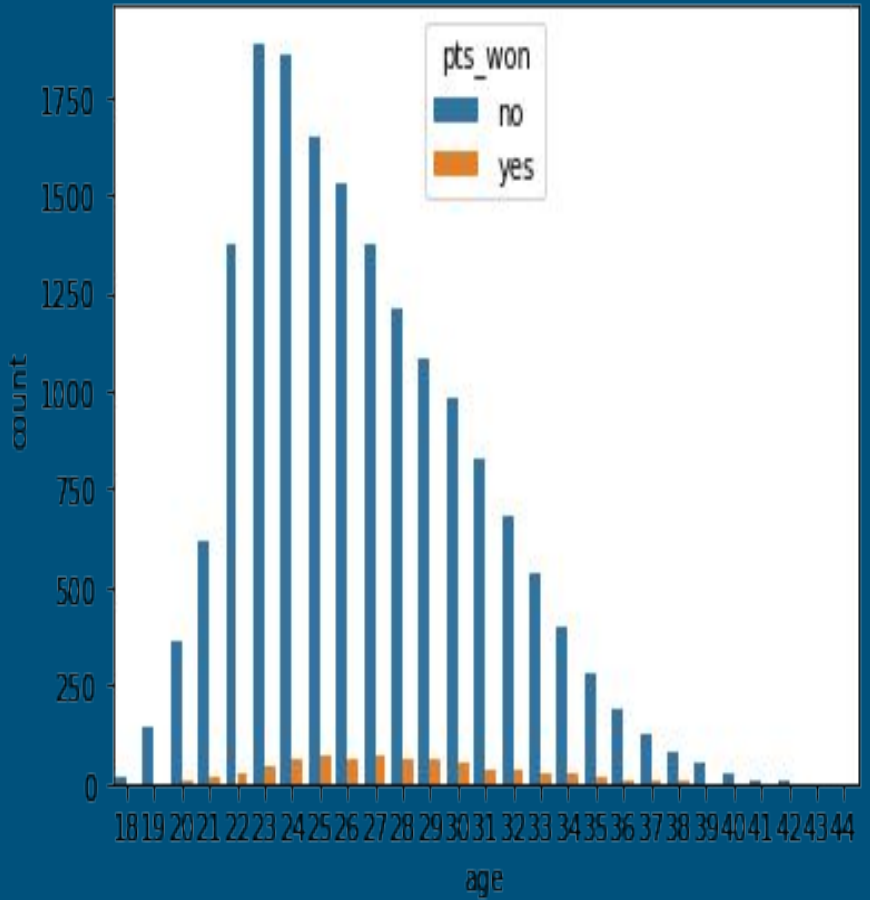
Results

To understand these results, I will generate an accuracy score and a confusion matrix. Understanding the confusion matrix of the results is vital.



Important insights





Important insights