Strike Prediction

Kyle Halkett and James Neri

Table of Contents

- Project Background
- Project Goals and Scope
- Solution Approach
- Project Plan
- Attribute Explanation
- Project Demonstration
- Issues, Risks, Mitigating Activities, and Mitigation Plan
- Project Outcome
- Highlighted Group - Group 3
- Future Opportunities
- Lessons Learned

Project Background

Client - Kevin Hickey

Location - Fairfield University

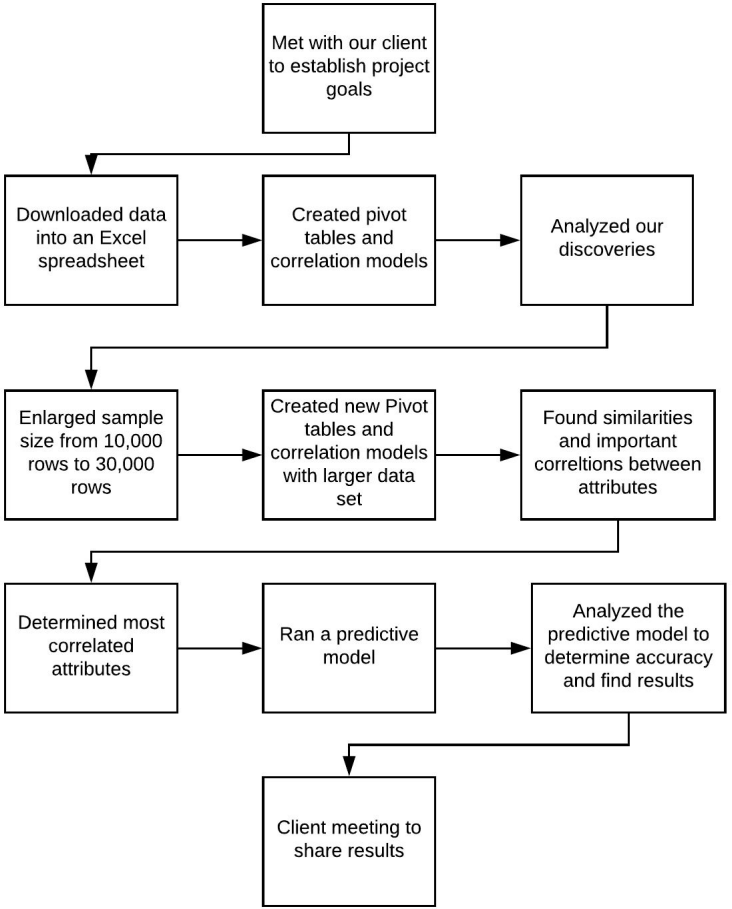
Our project:

- We used the website Baseballsavant.com to gather Statistical data from a Statcast Database, which is used to track the stats of MLB pitchers.
- With this statistical data we created a baseline of over 30,000 rows of attribute data, which allowed us to find correlations and run a predictive analytics model.

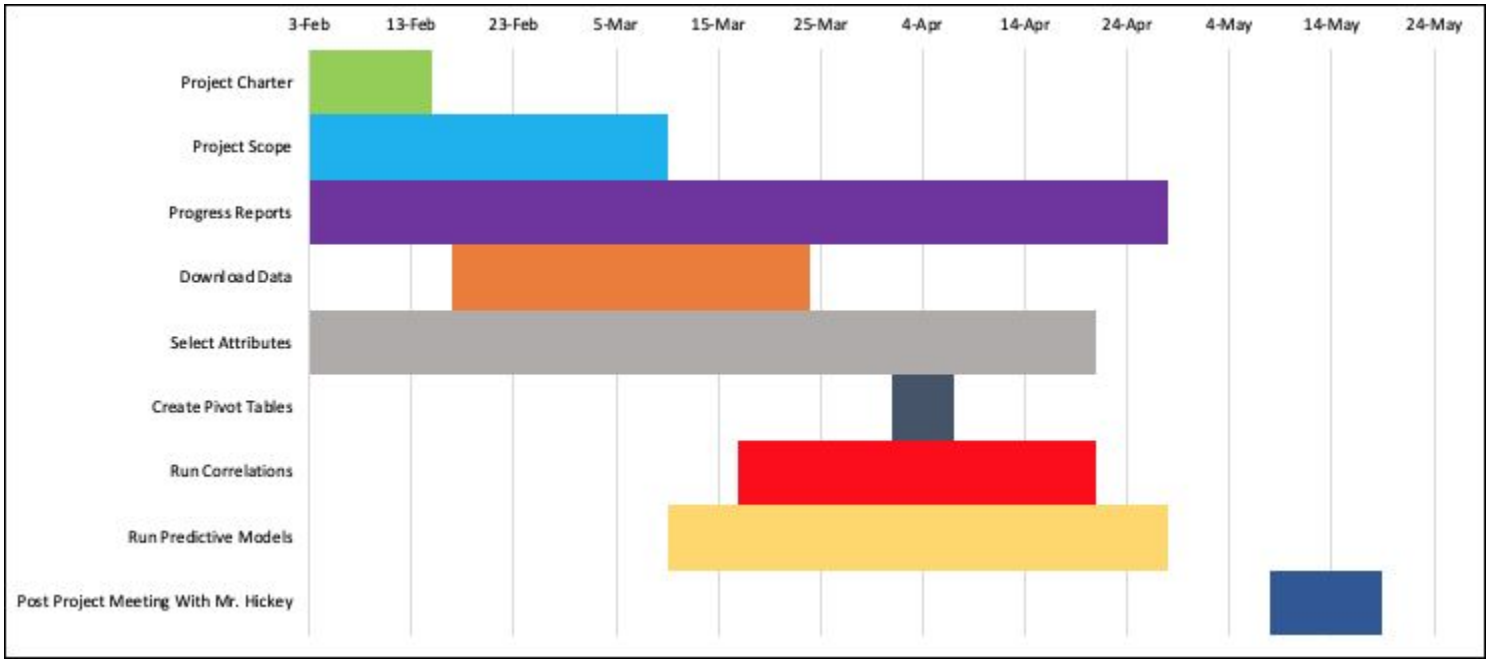
Project Goals and Scope

- Our client gave us free range to analyze the statistics we found to be most important in determining the result of a pitch.
- Defining the most important attributes allowed us to create various pivot tables, correlation models, and eventually run a predictive model.
- We will then upload our findings via an online database called GitHub, which will allow our client to use our baseline and predictive model to carry on his project.

Solution Approach



Project Plan



Attribute Explanation

Issues, Risks, Mitigating Activities, and Mitigation Plan

Issues

- Finding the right programs to run our data
- Having to restart our project whenever we added more data
- Rapidminer would unexpectedly quit which forced us to rely more heavily on Excel
- Increased program run time due to size of dataset

Mitigating Activities

- Chose RapidMiner because of previous class experience
- Pre-determined the size of the dataset to avoid this
- Determined the tasks we should perform in excel
- Chose one modeling method that was consistently the highest accuracy

Risks

- Increasing sample size could change the results of our predictive model
- Computer crashing from too much data
- Need for more advanced technical skills

Mitigation Plan

- Need to determine the amount of data needed to be most accurate
- Depending on the eventual size of this data set, special computers might be needed
- Hire outside analytics professionals

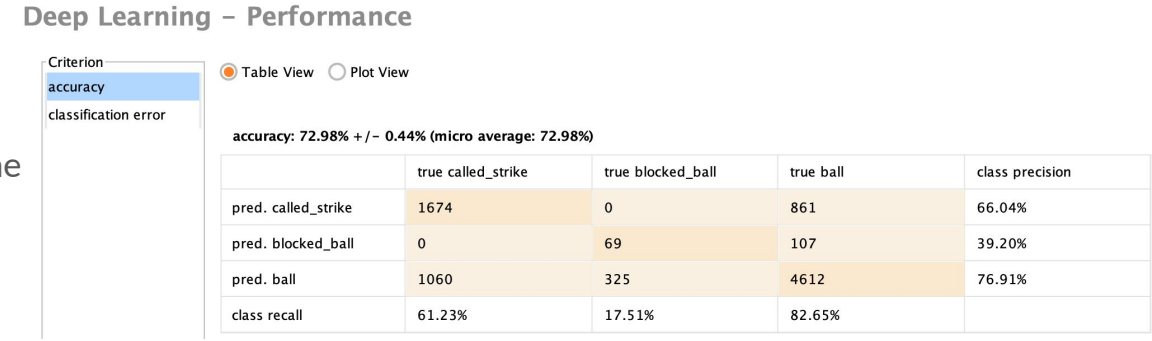
Project Outcome

- Created a baseline data set for Mr. Hickey to use in his further study of pitch prediction
    - Will use videos of pitches as well
- Explored different attributes to see which were most effective in predicting the result of a pitch
    - Identified key players in our models

Highlighted Group - Group 3

- Strong correlations:
  - **pfx_z** and **vy0** (-0.697)
  - **pfx_x** and **vx0** (-0.575)
  - **pfx_z** and **vz0** (-0.468)
- Deep Learning Model:
  - High predictive accuracy
  - **72.98%**
- Can predict a strike **66.04%** of the time

Attribu...	pfx_x	pfx_z	vx0	vy0	vz0
pfx_x	1	-0.188	-0.575	0.195	0.127
pfx_z	-0.188	1	0.039	-0.697	-0.468
vx0	-0.575	0.039	1	-0.135	-0.102
vy0	0.195	-0.697	-0.135	1	0.448
vz0	0.127	-0.468	-0.102	0.448	1



Future Opportunities

- Gather more data from the Statcast database in order to increase sample size
- Compare hard statistics with imaging analysis
- Baseball pitch prediction could become more accurate in a shorter amount of time
    - Ball or strike before it leaves the pitchers hand

Lessons Learned

- Time management
  - Take into account the little things that you would forget about
- Communication skills
  - Needed more frequent contact with our client
- Staying fresh on our technical skills
  - RapidMiner
  - Excel

Questions?