This write up explains my thought process while finding what and how variables contribute to a pitcher's fastball velocity. I will be discussing the assumptions, data exploration, preprocessing and the methods used to select these variables.

Assumptions

These are the assumptions that I am making while developing the model,

- Without knowing the variable names, consider all the biomech variables as equals. Subject to change if more context about these variables is provided.
- Correlation check between variables and velocity has been done while assuming the absolute value of Pearson Coefficient greater than 0.7 is highly correlated and anything less than or equal to 0.7 is safe.
- Alpha of 0.3 for Lasso Regularization is a good cut off parameter with a scaled dataset.

Data Exploration & Preprocessing

The data provided consists of 711 fastballs each thrown by unique players referenced by their ids. The columns consist of variables like handedness, height & weight of the pitchers, velocity of the fastball, and finally, 228 unnamed variables which could relate to multiple biomechanical variables, like Knee Extension, Hip-Shoulder Separation, etc. or even pitch characteristics, like Release Height, Extension, etc. that pitchers may use to throw hard.

To start off with the data exploration, I found that there were 26 NaN values each in three columns, *Var_214*, *Var_215* and *Var_216*. To solve this, I checked the correlation between the variables and the velocity to understand if they have any major effect on it. With their correlation scores around 0, the complete column was removed. I didn't want to remove the 26 rows, as there were already only 711 rows. After that I focused on the handedness of the player and noticed the imbalance between right handed (72.4%) and left handed (27.5%) pitchers. This can affect the variables being recorded within the unnamed columns, as the axis may be different for either of them. We cannot just drop the *Hand* column, and to take the variable in account, I used One-Hot encoding to code 'R' as 1 and 'L' as 0. Outliers within the dataset will be discussed after feature selection, since there are a lot of variables to just look at without any selection.

Methodology

The methodology section discusses the feature selection techniques employed, the variables selected, the clusters made and the models developed. Note that, if given more context about the data and the variables selected, it would lead to a more refined conclusion. I have done my best to make the assumptions required to come to an answer.

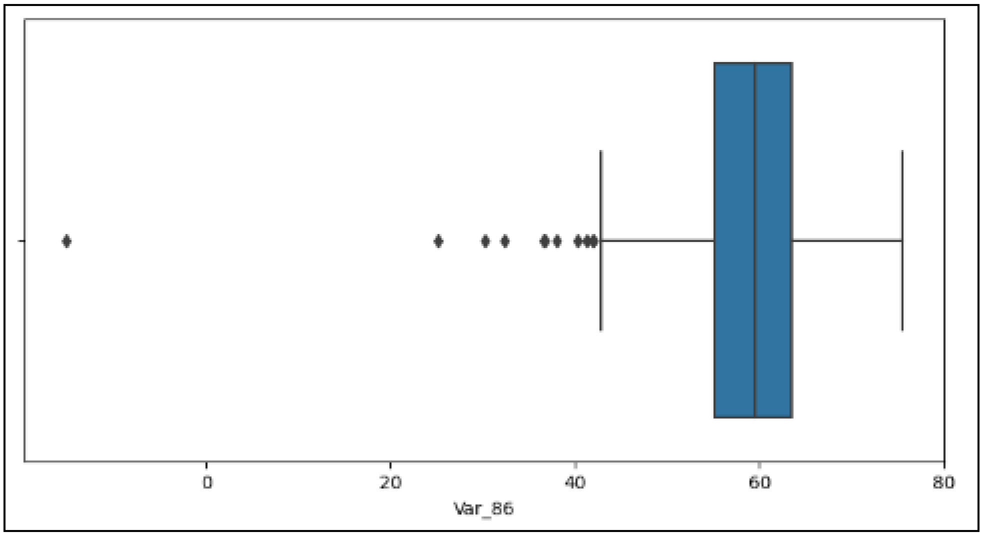
Feature Selection

Since we have 233 columns and only 711 rows, we would need to focus on selecting those variables that may have an impact over the target variable. For this, I used the two prominent feature selection methods for comparatively smaller datasets, Correlation Coefficient and Lasso Regression.

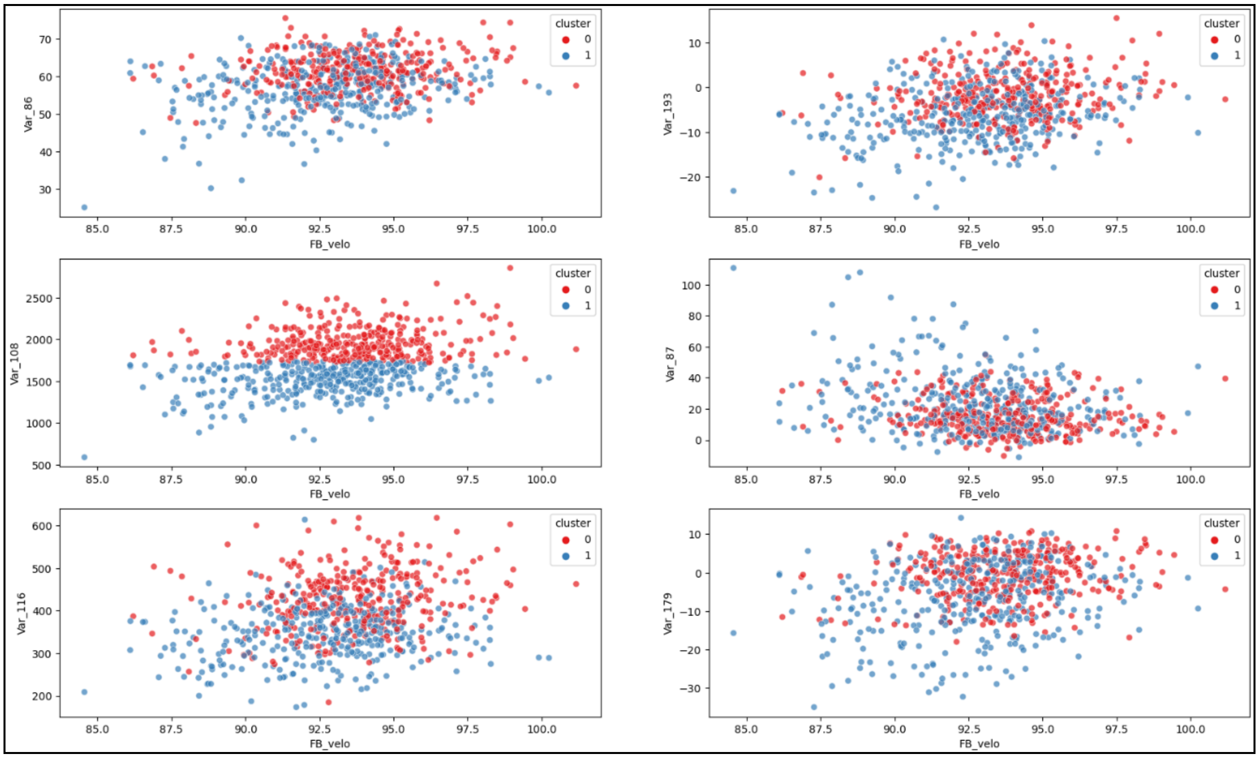
Correlation Coefficient

For this technique, I calculated the correlation coefficient between the variables in the dataset and the velocity column using Pearson Correlation Coefficients. Since we want to focus on a set of variables to find out how a pitcher uses them to increase their fastball velocity, we get the top 20 variables with the highest absolute correlation score. We then check for multicollinearity between the selected variables and get rid of the variables that are correlated to each other except the one that has the highest absolute correlation with the velocity amongst the variables being removed.

Using this method, I got 7 variables *Var_86*, *one_hot_hand* (the dummy variable for handedness), *Var_193*, *Var_108*, *Var_87*, *Var_116*, and *Var_179*. These variables were then checked for outliers, out of which just Var_86 had an outlier that stood out of the spread of data as depicted from the box plot. Even though some of the other selected variables had outliers, I have decided to include them till I can get more context about what the variables stand for and also that there are just 711 rows.



We can see from the plot below that *cluster_1* (blue) has values that are mostly related to lower fastball velocity, while the other cluster has values with comparatively higher *FB_velo*.

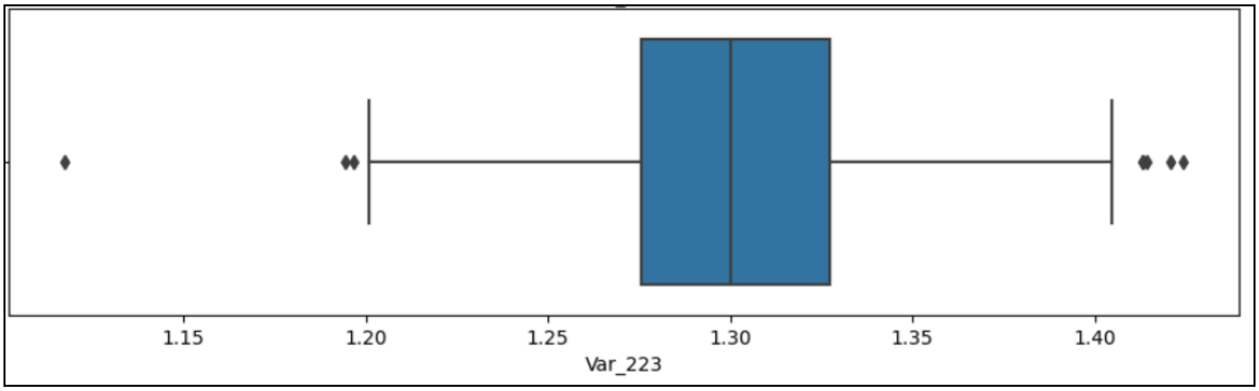


This data was then used to create a simple linear regression model and tree-based xgboost model to calculate the root mean square error (for how well the model is doing), RMSE, and the R^2 score (how much variance is explained by the variables). I got an R^2 score of 0.165 from the linear regression model and an RMSE of 2.008.

Lasso Regression

The second feature selection method I implemented was using Lasso Regression on all the variables in the dataset except the *player_id* and *FB_Velo*. With an alpha of 0.3, I got 15 variables whose coefficients weren't turned to 0 by the method. I checked for multicollinearity amongst the chosen variables, but there was only one pair of variables that was above the assumed correlation threshold, *Var_86* & *Var_70*. Since, Var_86 has a higher correlation with velocity, I removed *Var_70* from the selected variables.

While checking for outliers, I removed the outlier standing out from the boxplot for variables like *Var_107*, *Var_181*, *Var_223* and *Var_225*. Again, there were other outliers that were present, but with no additional context, I decided not to remove them.



After taking care of the outliers, I decided to implement the same modeling methods for variables selected from Lasso Regression. I ended up getting an R^2 score of 0.413 and an RMSE of 1.723 from the linear regression and xgboost model.

Lasso Regression

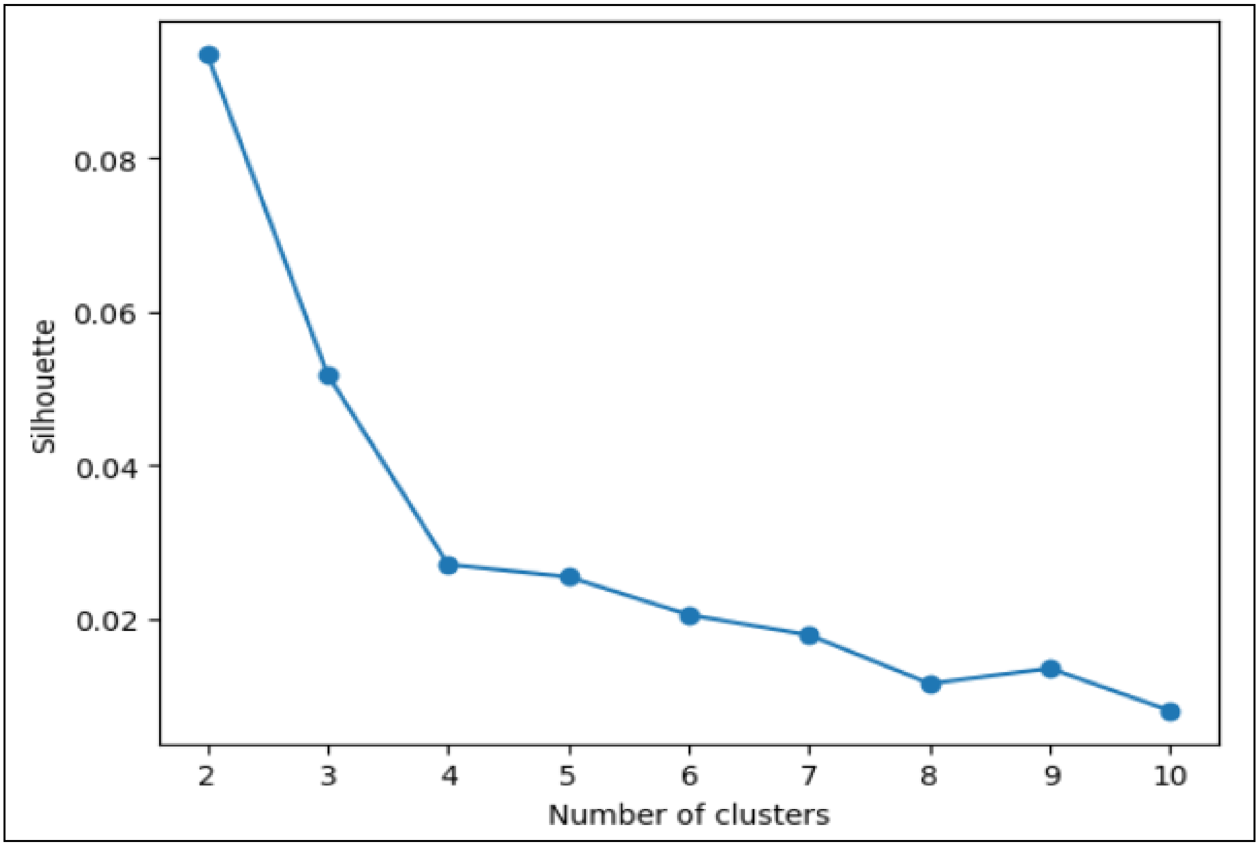
Since Lasso Regression/Regularization shrinks the coefficients of unimportant variables, I chose the variables selected from Lasso Regression. Also, the RMSE of the lasso model was lower than the alternative.

Model

I used XGBoost Regressor to model the selected variables, and use SHAP values to understand the importance value of each feature. SHAP allows for a deeper dive into how the variables in the model contribute towards the target variable, in this case, velocity. I will use the SHAP values after clustering the data on the selected variables.

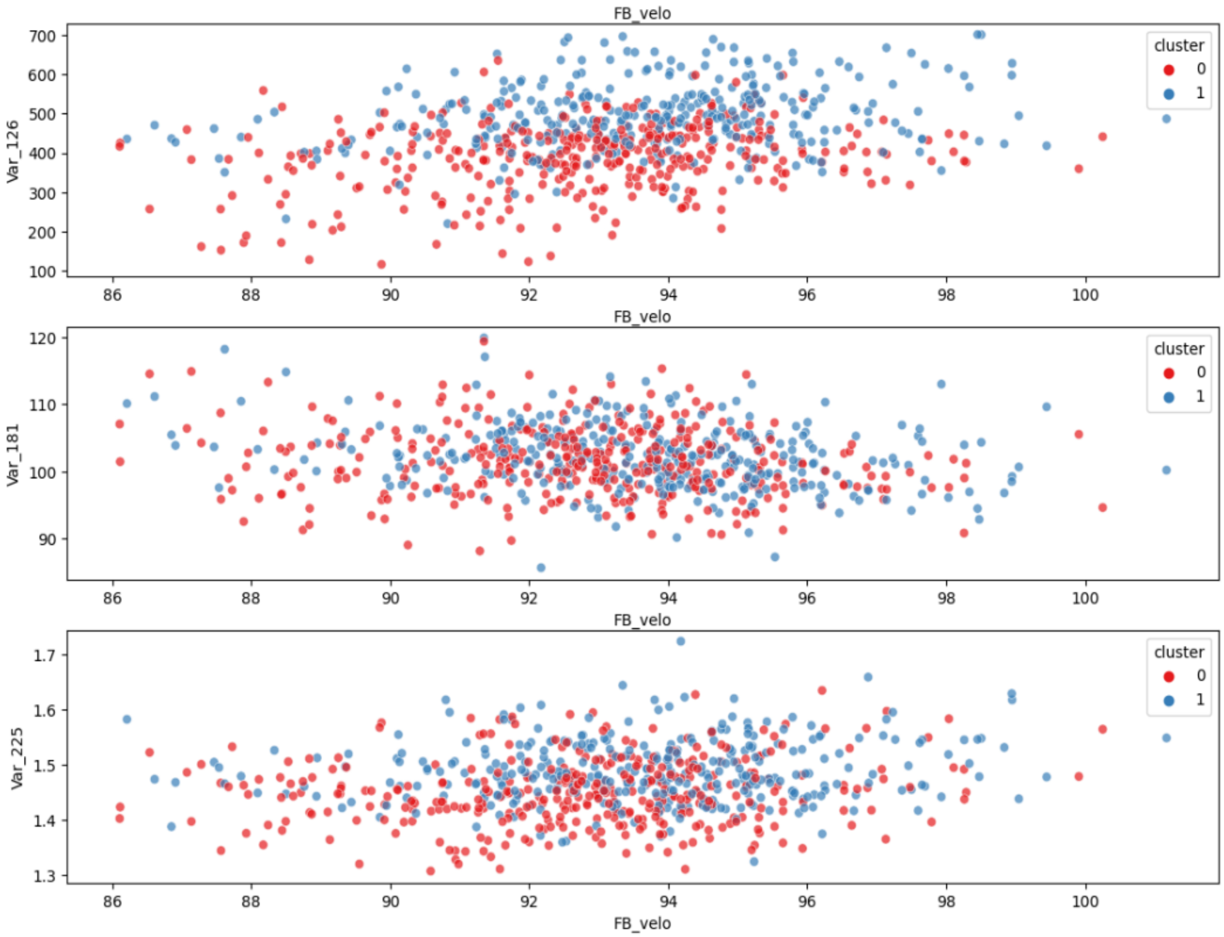
Clustering (KMeans)

Clustering was implemented using the KMeans algorithm, and the optimal clusters were chosen using the Silhouette analysis. This method of identifying the optimal number of clusters studies the separation distance between the clusters and has an advantage over the elbow method to find outliers present in the clusters. We chose the number of clusters as 2, since it had the max silhouette score at that cluster.



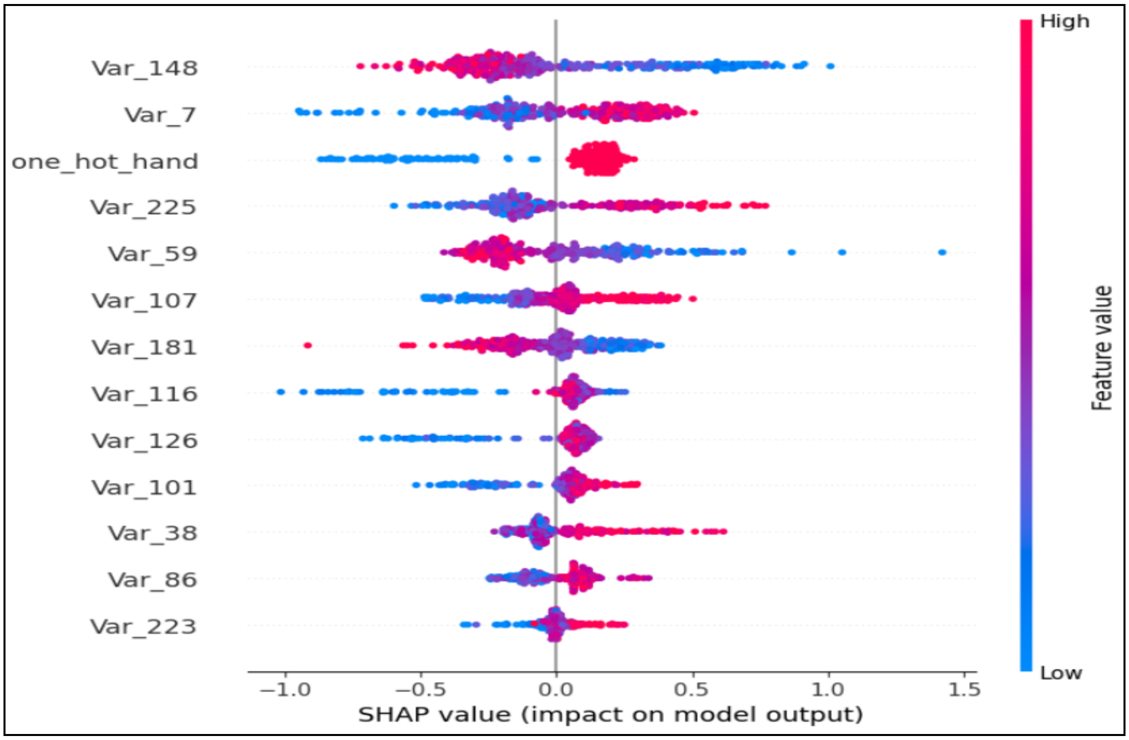
While there is not a clear correlation visible from the graphs, we can see from plotting the clusters over *Var_126*, *Var_181* and *Var_225* vs *FB_velo* that cluster_0 is broadly associated with lower velocity irrespective of the value of the variable. After checking the mean velocity for both clusters, it is conclusive that *cluster_0* is associated with lower velocity with an average of 92.79 and *cluster_1* is associated with higher velocity with an average of 93.64. Obviously, with more data I believe we can make this even more conclusive than it is right now.

Now, I will focus on individual clusters and look at how the variables are affecting the velocity inside those clusters. We will look at waterfall plots for a high and a low velocity pitch as well as a summary plot for each of the two clusters to clearly identify which variables may lead to a lower or a higher velocity.

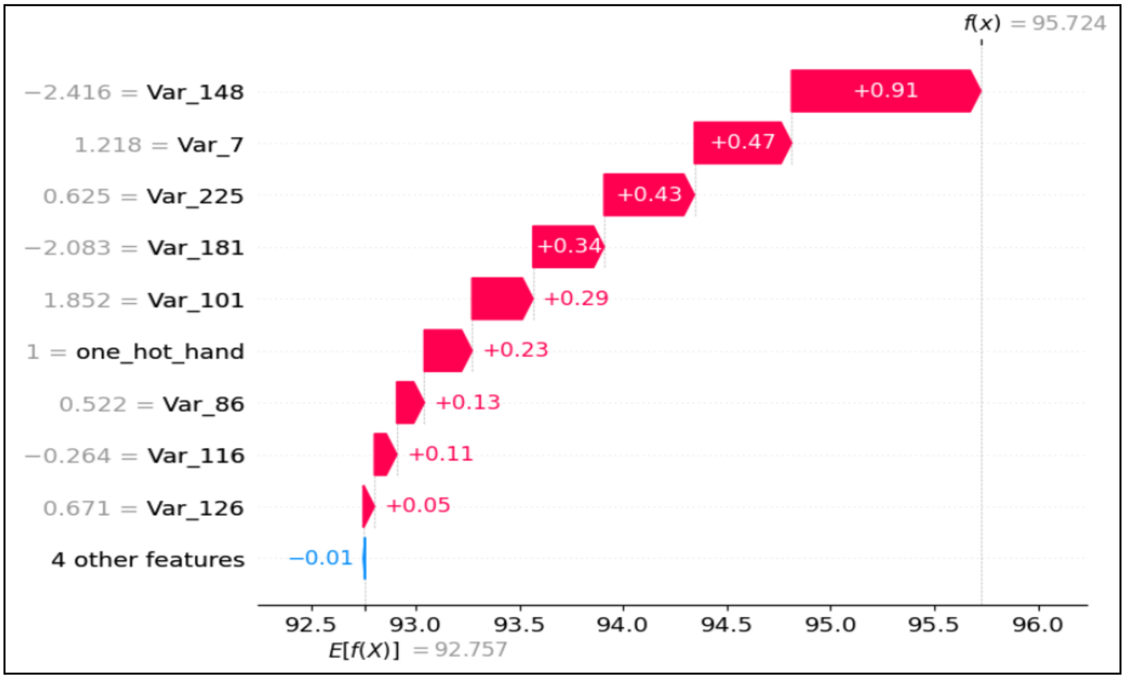


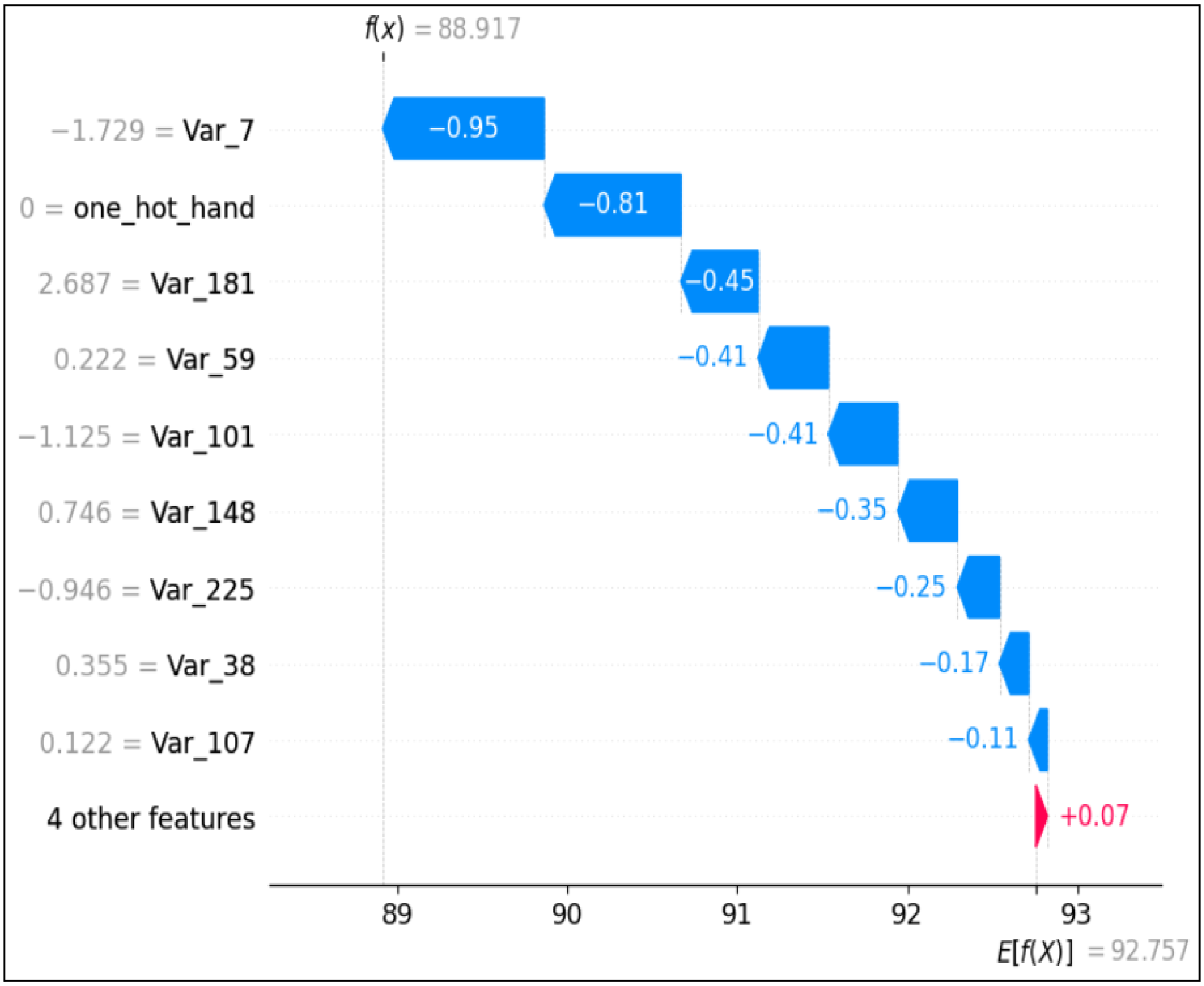
Cluster 0

For the summary plot for cluster 0, we see that the three most important features affecting the velocity are *Var_148*, *Var_7* and *one_hot_hand* (which is a binary variable). In the case of *Var_148*, we see a negative impact on the model for higher values of *Var_148*, while a more positive impact on the model for lower values of *Var_148*. However, in case of *Var_7*, we see a comparatively direct correlation, where if the value of *Var_7* is low, then the velocity will also be low and vice versa. Similarly, for *one_hot_hand*, when the value is 0 (or the pitcher is left handed), we see a lower model output, i.e., *FB_velo*, while for righty pitchers, the velocity is higher according to the summary plot. Again, it would be super beneficial to get more data, as we had quite a lot of data for righty pitchers than lefty ones. For the unknown variables, it will be quite fruitful to have some context about it to refine our decisions from the plots.
We can even count *Var_225* and *Var_59* as having an impact on the velocity.



While comparing the waterfall plot of two very different pitches predicted by the model, we see that *Var_7* is super insightful in differentiating between a high and a low velocity pitch, where a more positive value of *Var_7* leads to a positive contribution in the velocity. We also see that, in this cluster, a more negative value of *Var_148* leads to more velocity of the pitch. one_hot_hand being 0, or for a lefty pitcher, had a significant contribution towards lower velocity of a pitch, while the same variable being 1, or for a righty pitcher, wasn't super important for a high velocity pitch.

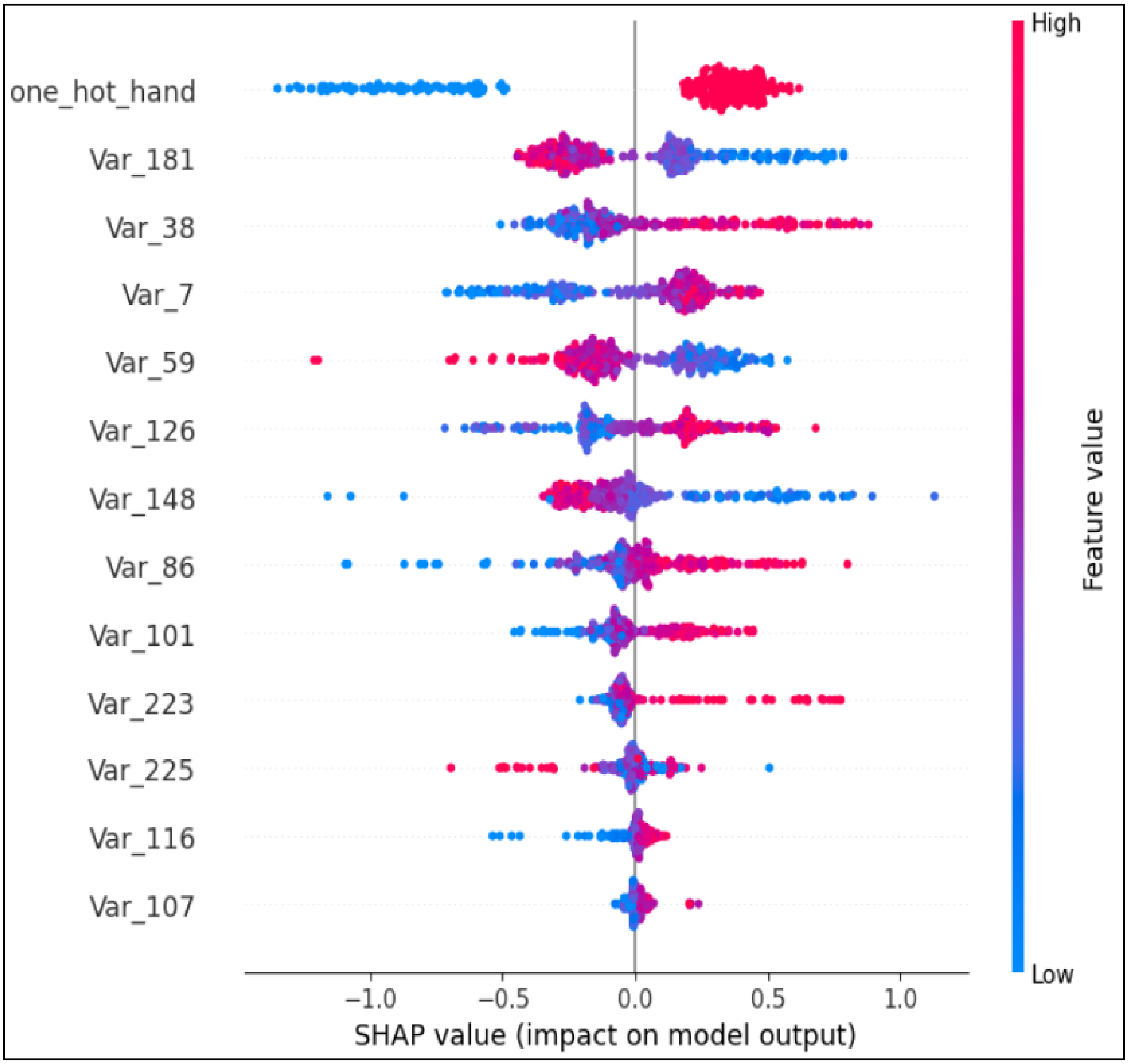




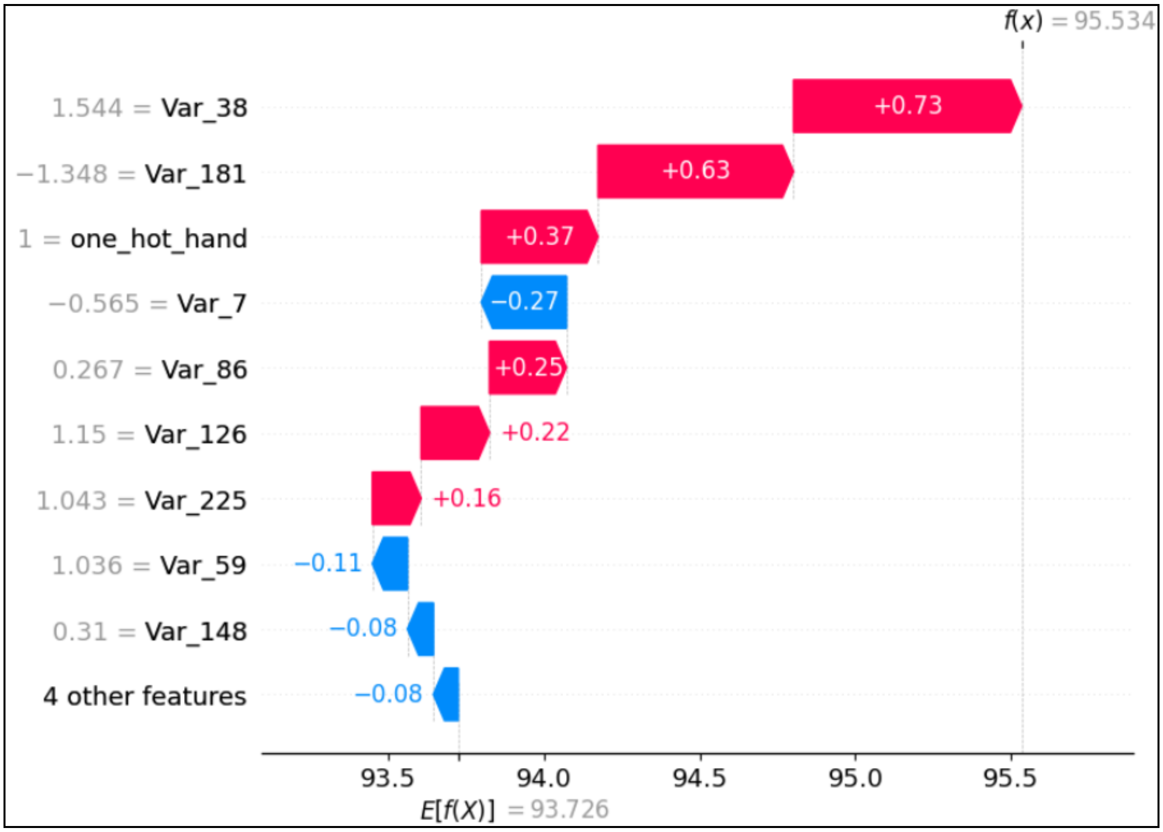
Cluster 1

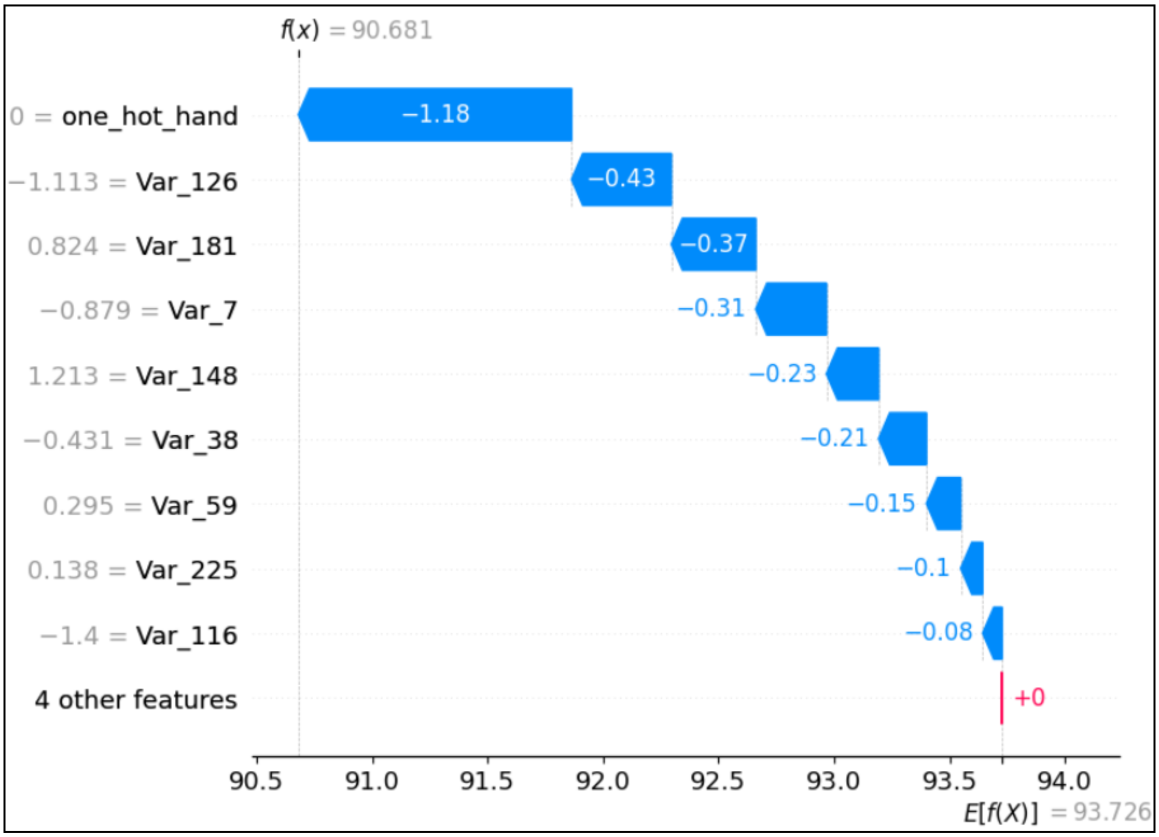
For the summary plot for *cluster_1*, we see that the top most important variables with a lot of contribution are *one_hot_hand*, *Var_181* and *Var_38*. For one_hot_hand, it is the same relationship as for *cluster_0*. If anything, I would say it is more defined in the case of *cluster_1*, where a lefty pitcher can have a very negative impact on velocity. *Var_181* has an inversely proportional relationship with the impact on velocity, similar to the previous cluster, however, it affects the velocity more strongly (being the 2nd most important feature). Finally, *Var_38* has a more direct impact on velocity, where an increase in *Var_38* feature value results in an increase in the velocity.

We can also keep *Var_7* and *Var_59* as variables that we can consider having an impact on velocity for this cluster.



We can confirm our analysis by looking at two pitches varying in velocity. For a righty (*one_hot_hand*=1) pitcher with a high *Var_38* and low *Var_181*, there is a positive impact on the expected value, thereby increasing the velocity. Similarly, for a lower velocity pitch, we see that it is from a lefty pitcher (*one_hot_hand*=0) with a relatively higher *Var_181* and lower *Var_38*.


f(x) = 95.534
1.544 = Var_38
−1.348 = Var_181
1 = one_hot_hand
−0.565 = Var_7
0.267 = Var_86
1.15 = Var_126
1.043 = Var_225
1.036 = Var_59
0.31 = Var_148
4 other features
+0.73
+0.63
+0.37
−0.27
+0.25
+0.22
+0.16
−0.11
−0.08
−0.08
93.5
94.0
94.5
95.0
95.5
E[f(X)] = 93.726


f(x) = 90.681
0 = one_hot_hand
−1.113 = Var_126
0.824 = Var_181
−0.879 = Var_7
1.213 = Var_148
−0.431 = Var_38
0.295 = Var_59
0.138 = Var_225
−1.4 = Var_116
4 other features
−1.18
−0.43
−0.37
−0.31
−0.23
−0.21
−0.15
−0.1
−0.08
+0
90.5
91.0
91.5
92.0
92.5
93.0
93.5
94.0
E[f(X)] = 93.726

Conclusion

In conclusion, the analysis provides comprehensive insights into the significant variables influencing a pitcher's fastball velocity. Based on the explored data, the handedness of the pitcher, represented by the *one_hot_hand*, appears to be a defining factor, with right-handed pitchers tending to exhibit higher velocities. Variables such as *Var_148*, *Var_7*, *Var_181*, and *Var_38* have also shown either a direct or inverse relationship with fastball velocity across the clusters. However, without specific context regarding the underlying nature of these variables, conclusions remain somewhat speculative. The disparity in data representation between right and left-handed pitchers suggests that gathering more balanced data, especially for lefty pitchers, can yield more definitive results. It's also imperative to understand the real-world implications of the unnamed variables to make actionable recommendations. The use of both correlation coefficients and Lasso Regression for feature selection has proven fruitful, with the latter being more effective in this dataset. For future work, understanding the biomechanical significance of the selected variables, coupled with a larger dataset, can lead to more precise predictions and understanding of the dynamics of fastball velocity.