Intro + Background: MLB Scouting Methodologies and Performance Prediction

MLB scouting departments often follow different methodologies around scouting. Whether to draft players out of high school or out of college is an important scouting decision. We want to predict the performance of a player based on NCAA statistics on how they'll perform in the MLB. A plethora of data is available surrounding baseball performance. Hitting metrics are often focused on for contribution to team performance, but there are also defensive statistics available, and pitchers can be compared. It's important to realize that a team's number of outs available. This is a scarce resource, so whichever action a player can do to reduce the depletion of this resource is good for the team. A team's goal should be to win games, and the best way to do this is to score runs. So the more runs a team can score, the more games they should win. [3].

Baseball is a sport rich with data, and because every team plays 162 games in a season, variance is lower with the larger sample size. Some popular hitting metrics include batting average (player hits / number of at bats), RBI (runs batted in), and home runs. OPS, "on-base plus slugging" is correlated with the number of runs scored, and is very popular in sabermetrics. It adds the on-base percentage and slugging percentage (total bases reached / at bats). These two metrics individually are also helpful to look at. Popular pitching metrics include wins (pitcher when a team takes the lead and doesn't lose it, with some exceptions), ERA (earned run average), and strikeouts. [1, 2]

Problem Definition

MLB teams face significant uncertainty when drafting players from college baseball. Despite having access to NCAA statistics, it's challenging to predict how well a college player will perform at the professional level. This creates a substantial risk in the player selection and development process.

Project Motivation

Scouting Decisions: MLB teams must make critical decisions between drafting players from high school or college. This requires better predictive tools to evaluate college talent. [4]

Business Impact: Reduce recruitment risks for MLB teams. Teams invest millions into draft picks, but predicting performance remains uncertain.

Analytical Opportunity: Machine learning can potentially find patterns in college performance that indicate future MLB success, making this a challenging and fun problem worth solving.

Methods

Preprocessing

Before training our models, we standardized our data using sklearn.preprocessing.StandardScaler. Standardizing data can be helpful for this case because we have many different features that are on different scales such as batting average and number of runs. Having features on the same scale ensures that features with a larger scale do not disproportionately influence the model. We also employed one-hot encoding using pandas.get_dummies to handle categorical variables in our dataset, allowing the models to properly interpret non-numeric features. To handle missing values that could affect our analysis, we utilized sklearn.impute.SimpleImputer with mean, ensuring our models had complete data to work with. Finally, we split our data into training (80%) and testing (20%) sets to properly evaluate model performance.

Model

For our model we selected a Random Forest Regressor from sklearn.ensemble. RandomForestRegressor. There are a few advantages that make this a good choice for our use case. First, because the model selects a random subset of features to be used for each decision tree, it has a lower chance of overfitting with a sufficient number of estimators. For this same reason, the model is well-suited to handle high-dimensional data, which we have in our use case. Additionally, Random Forests make it easy to evaluate the contribution of features for the model. This is beneficial to allow us to determine which statistics are most predictive of a player's professional batting average.

We also implemented Linear Regression from sklearn.linear_model.LinearRegression as our second model. Its simplicity and interpretability make it a strong choice for this baseball prediction task. The linear model can capture direct relationships between college and professional statistics and tell us how each college statistic contributes to MLB performance through its coefficients, providing clear insights into which metrics are most important for predicting future success. For our third model, we employed Support Vector Regression with an RBF kernel from sklearn.svm.SVR. SVR was chosen to explore potential non-linear relationships between college and professional statistics that simpler models might miss. The RBF kernel allows the model to capture more complex patterns in player development trajectories, though this added complexity didn't necessarily translate to better predictions in our case

Results + Discussion

After implementing and evaluating three distinct machine learning models for predicting MLB batting averages from NCAA statistics, we achieved varying levels of success with each approach. Here we present our findings and analysis:

Quantitative Performance Metrics [2]:

Random Forest Regressor:

MSE: 0.000377

R-squared: 0.973

Linear Regression:

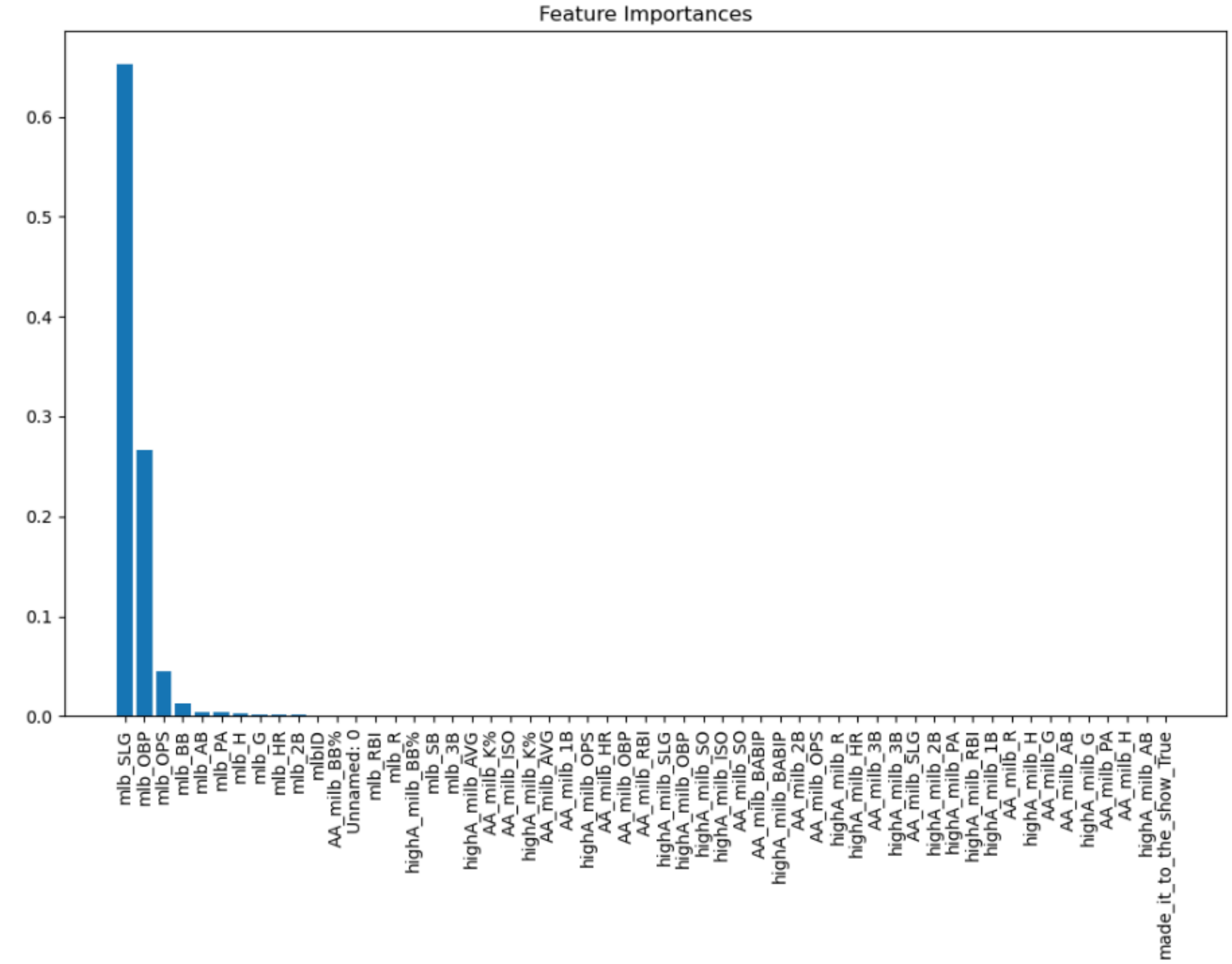
MSE: 0.00190

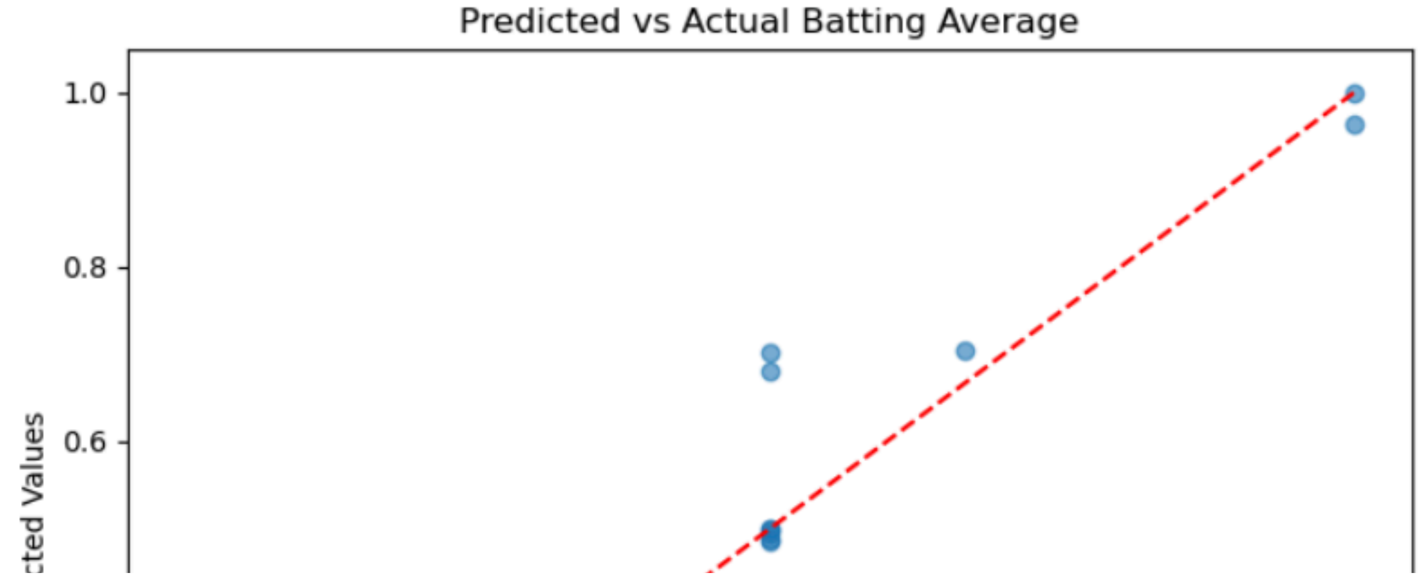
R-squared: 0.866

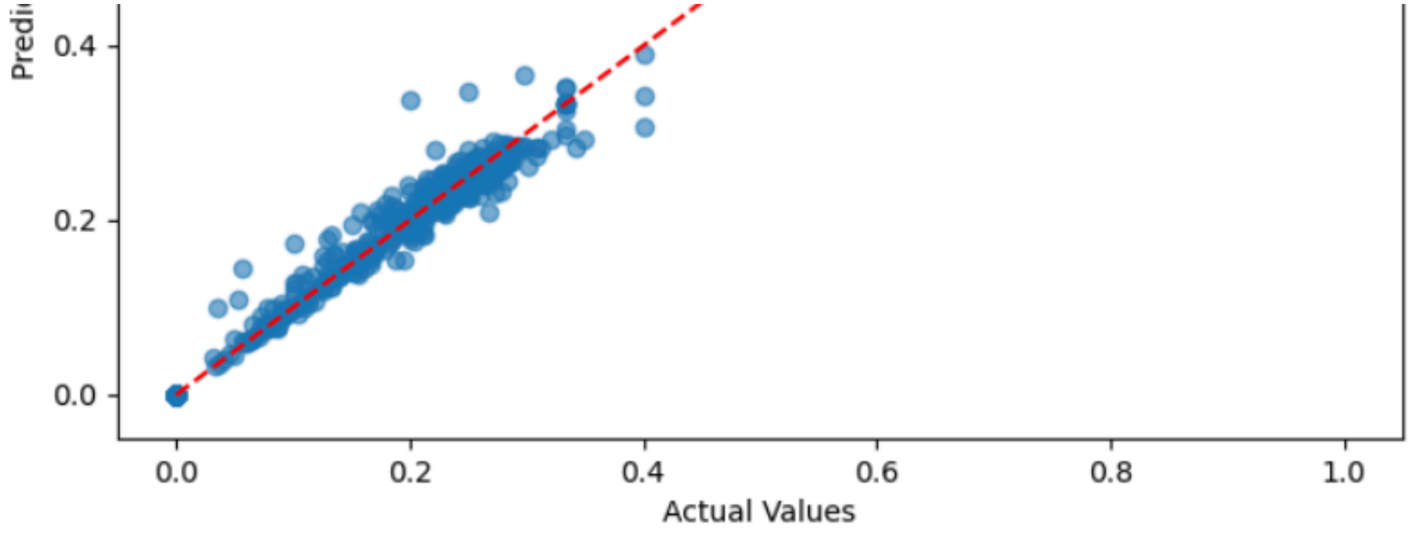
Support Vector Regression:

MSE: 0.00266

R-squared: 0.812


Feature Importances
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mlb_SLG
mlb_OBP
mlb_OPS
mlb_BB
mlb_AB
mlb_PA
mlb_H
mlb_G
mlb_HR
mlb_2B
mlbID
AA_milb_BB%
Unnamed: 0
mlb_RBI
mlb_R
highA_milb_BB%
mlb_SB
mlb_3B
highA_milb_AVG
AA_milb_K%
AA_milb_ISO
highA_milb_K%
AA_milb_AVG
AA_milb_1B
highA_milb_OPS
AA_milb_HR
AA_milb_OBP
AA_milb_RBI
highA_milb_SLG
highA_milb_OBP
highA_milb_SO
highA_milb_ISO
AA_milb_SO
AA_milb_BABIP
highA_milb_BABIP
AA_milb_2B
AA_milb_OPS
highA_milb_R
highA_milb_HR
AA_milb_3B
highA_milb_3B
AA_milb_SLG
highA_milb_2B
highA_milb_PA
highA_milb_RBI
highA_milb_1B
AA_milb_R
highA_milb_H
AA_milb_G
AA_milb_AB
highA_milb_G
AA_milb_PA
AA_milb_H
highA_milb_AB
made_it_to_the_show_True

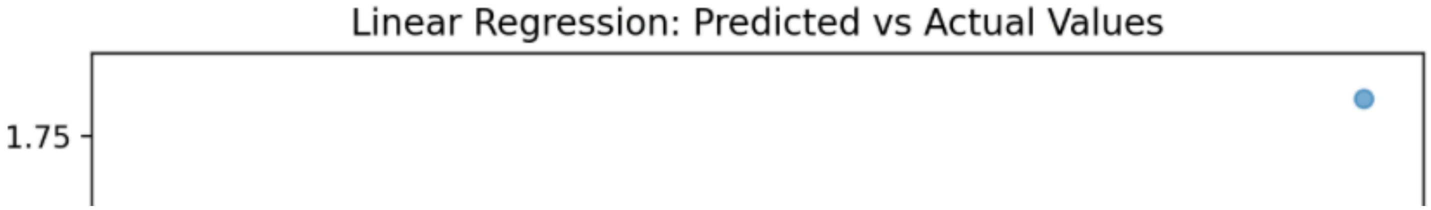

Predicted vs Actual Batting Average
1.0
0.8
0.6
cted Values

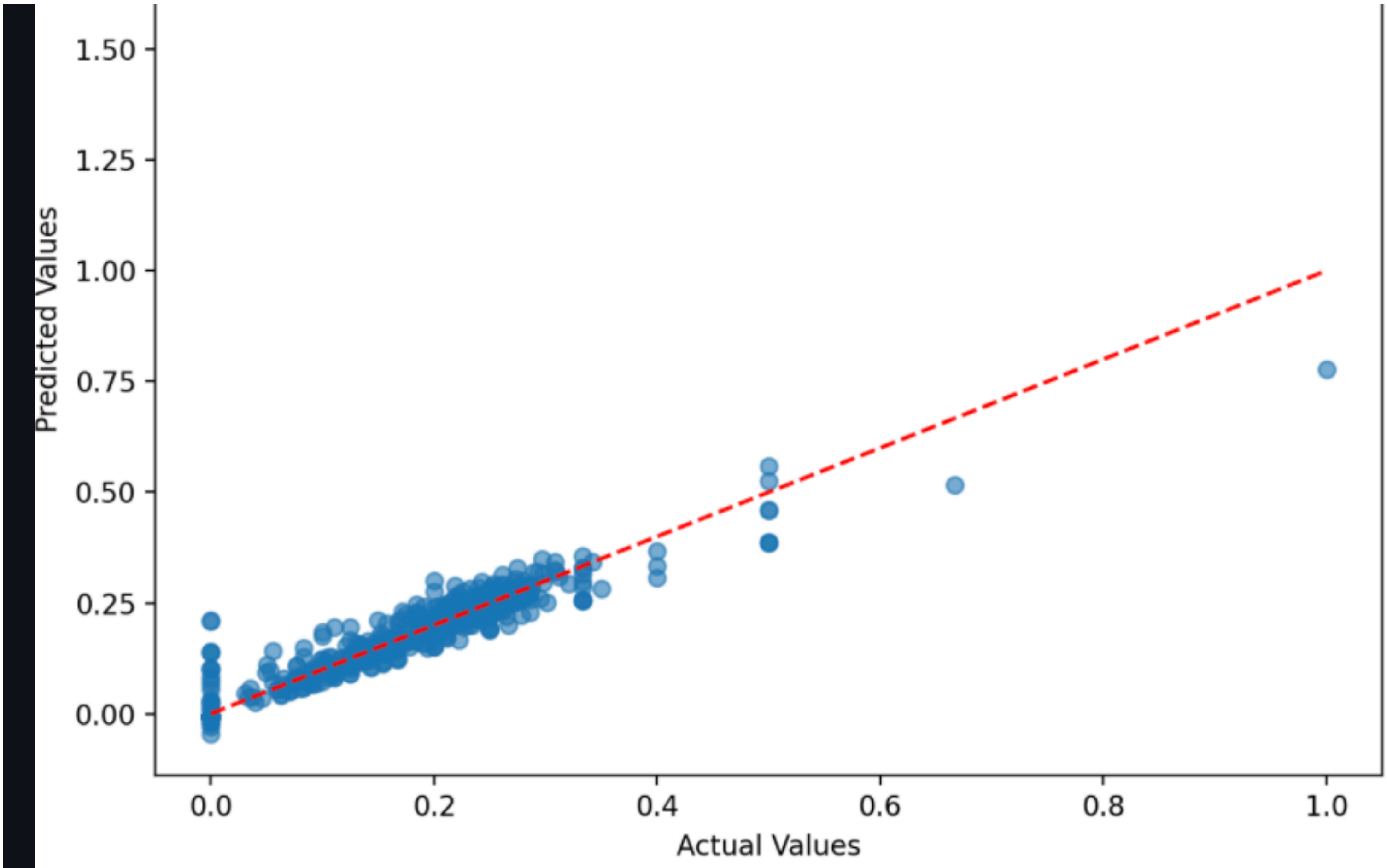


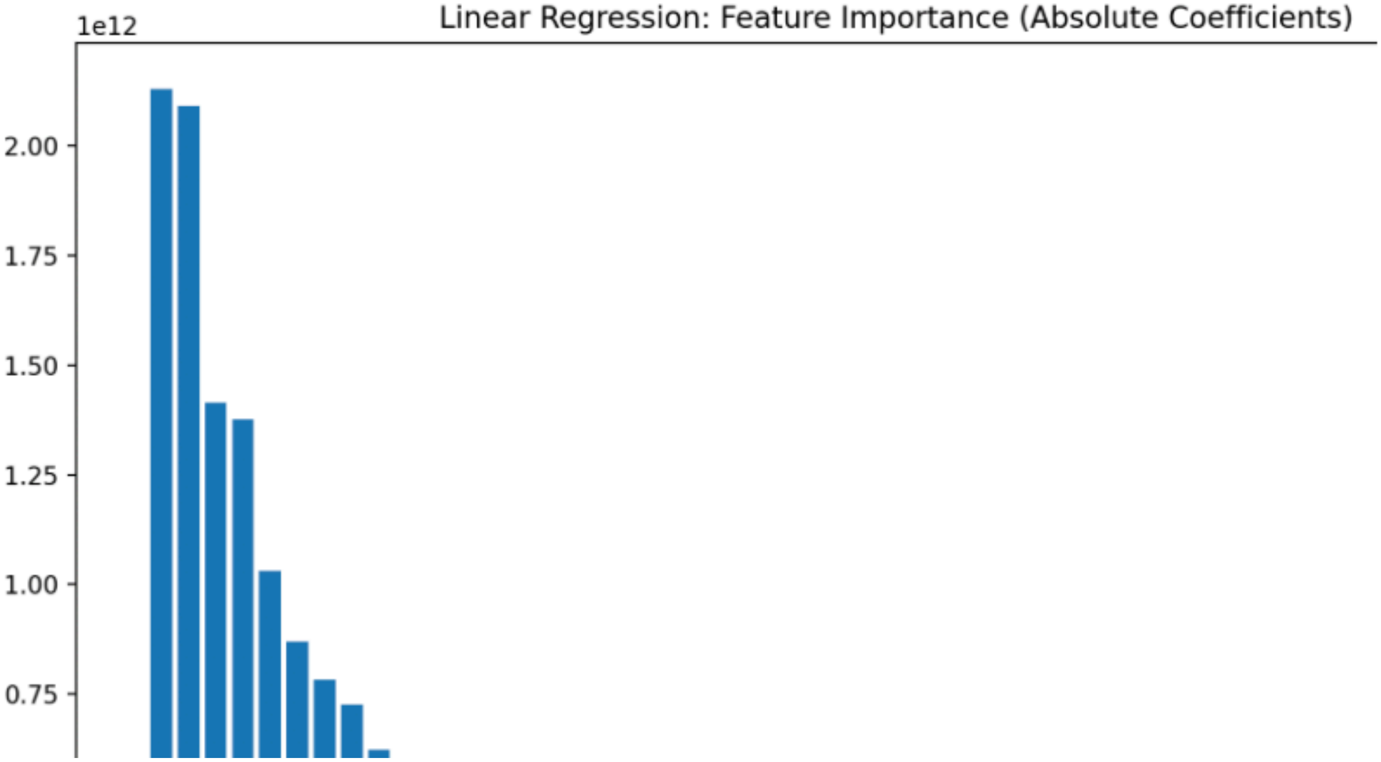
These metrics indicate that the model has a high predictive accuracy. Specifically: The low MSE shows minimal prediction error, meaning the predicted batting averages are close to the actual values. The R-squared value of 0.974 implies that the model explains 97.4% of the variance in batting average outcomes This is a strong indication of a well-fitting model. The high R^2 suggests that the model captures the majority of necessary information The Random Forest Regressor performed well because it effectively handles high-dimensional data by using random subsets of features for each tree, reducing overfitting and improving generalization. Key features like mlb_SLG and mlb_OBP were identified as strong predictors, aligning with known baseball statistics [1]. With 100 estimators, the model balances accuracy and generalization, achieving a high R^2 and low MSE.

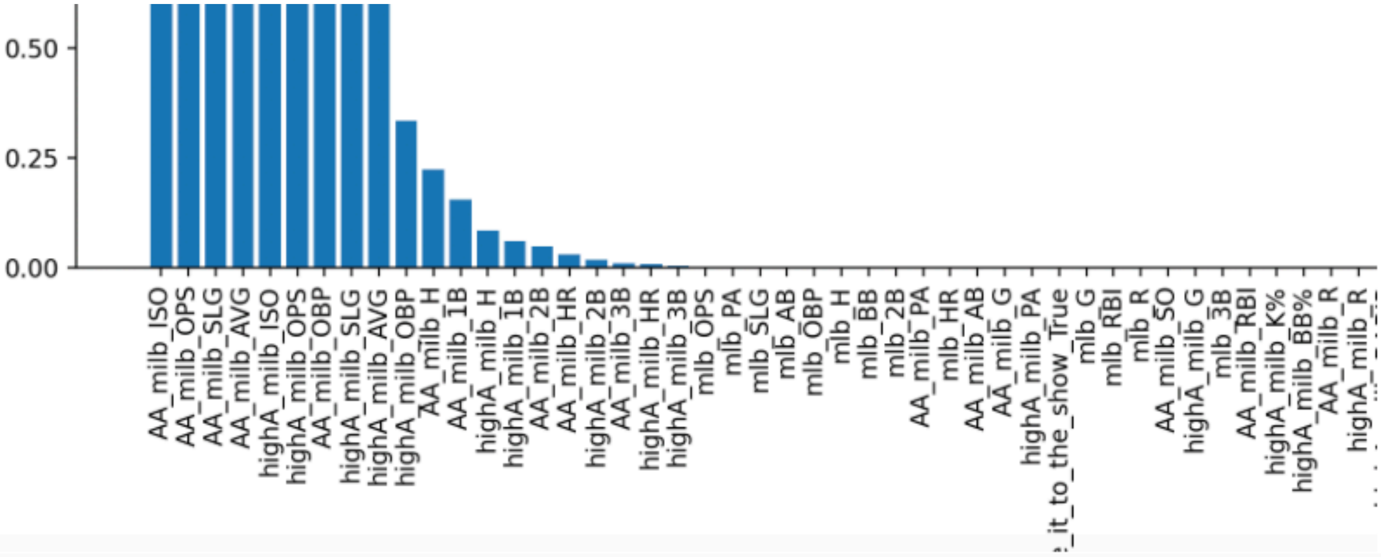
The **Random Forest Regressor** performed well because it effectively handles high-dimensional data by using random subsets of features for each tree, which helps reduce overfitting and improves generalization. Key features, such as **mlb_SLG** and **mlb_OBP**, were identified as strong predictors, aligning with known baseball statistics. With 100 estimators, the model balances accuracy and generalization, achieving both a high R^2 and low MSE.

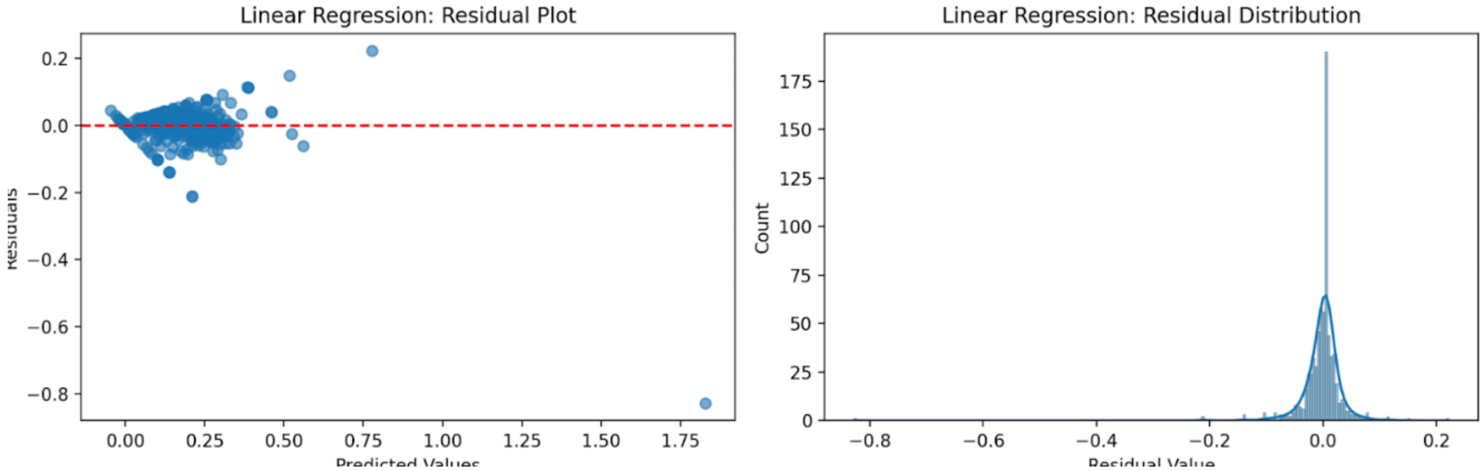
Linear Regression

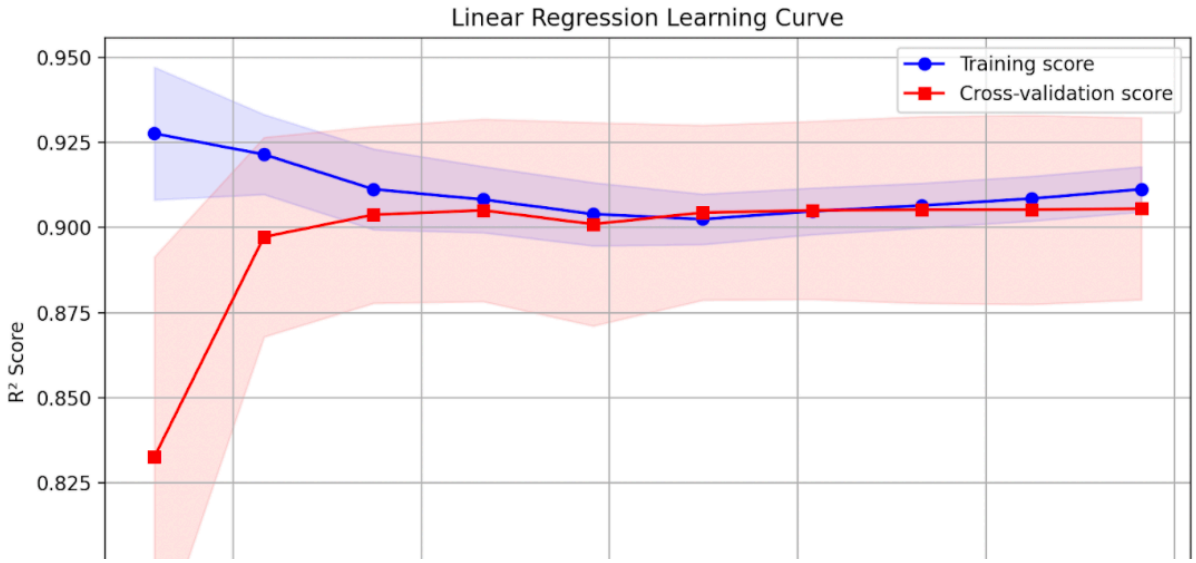


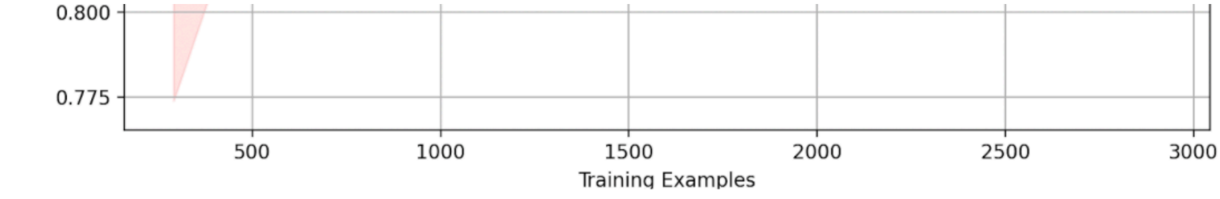

Predicted Values
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Actual Values


Linear Regression: Feature Importance (Absolute Coefficients)
1e12
2.00
1.75
1.50
1.25
1.00
0.75


0.50
0.25
0.00
AA_milb_ISO
AA_milb_OPS
AA_milb_SLG
AA_milb_AVG
highA_milb_ISO
highA_milb_OPS
AA_milb_OBP
highA_milb_SLG
highA_milb_AVG
highA_milb_OBP
AA_milb_H
AA_milb_1B
highA_milb_H
highA_milb_1B
AA_milb_2B
AA_milb_HR
highA_milb_2B
AA_milb_3B
highA_milb_HR
highA_milb_3B
mlb_OPS
mlb_PA
mlb_SLG
mlb_AB
mlb_OBP
mlb_H
mlb_BB
mlb_2B
AA_milb_PA
mlb_HR
AA_milb_AB
AA_milb_G
highA_milb_PA
_it_to_the_show_True
mlb_G
mlb_RBI
mlb_R
AA_milb_SO
highA_milb_G
mlb_3B
AA_milb_RBI
highA_milb_K%
highA_milb_BB%
AA_milb_R
highA_milb_R


Linear Regression: Residual Plot
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Residuals
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Predicted Values
Linear Regression: Residual Distribution
175
150
125
100
75
50
25
0
Count
−0.8
−0.6
−0.4
−0.2
0.0
0.2
Residual Value

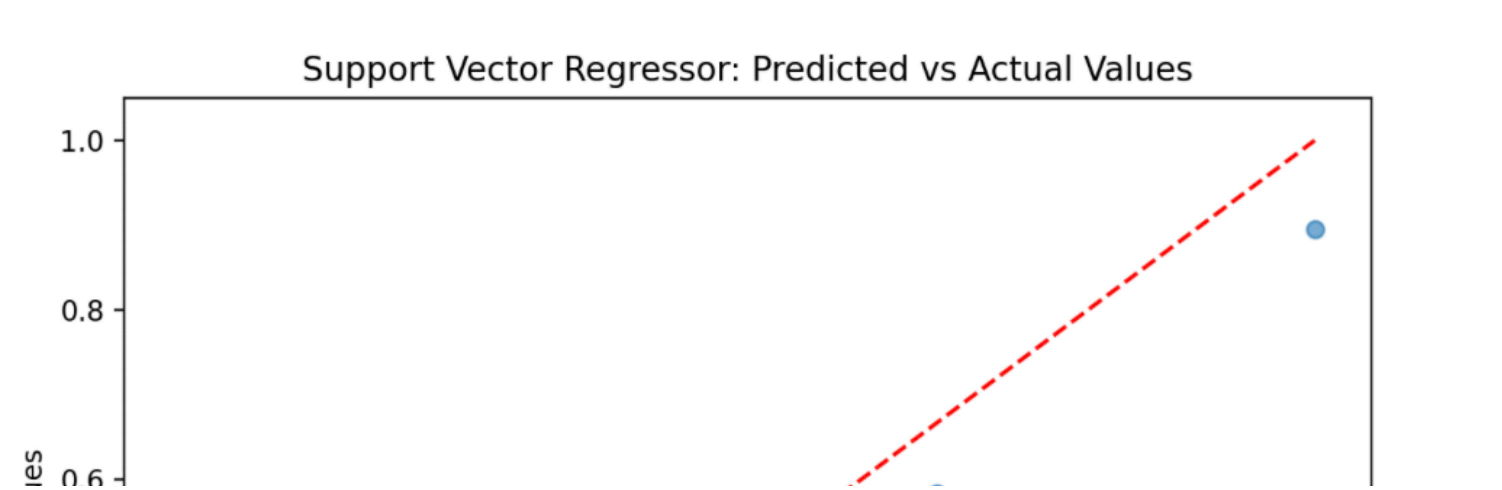

Linear Regression Learning Curve
Training score
Cross-validation score
0.950
0.925
0.900
0.875
0.850
0.825
R² Score

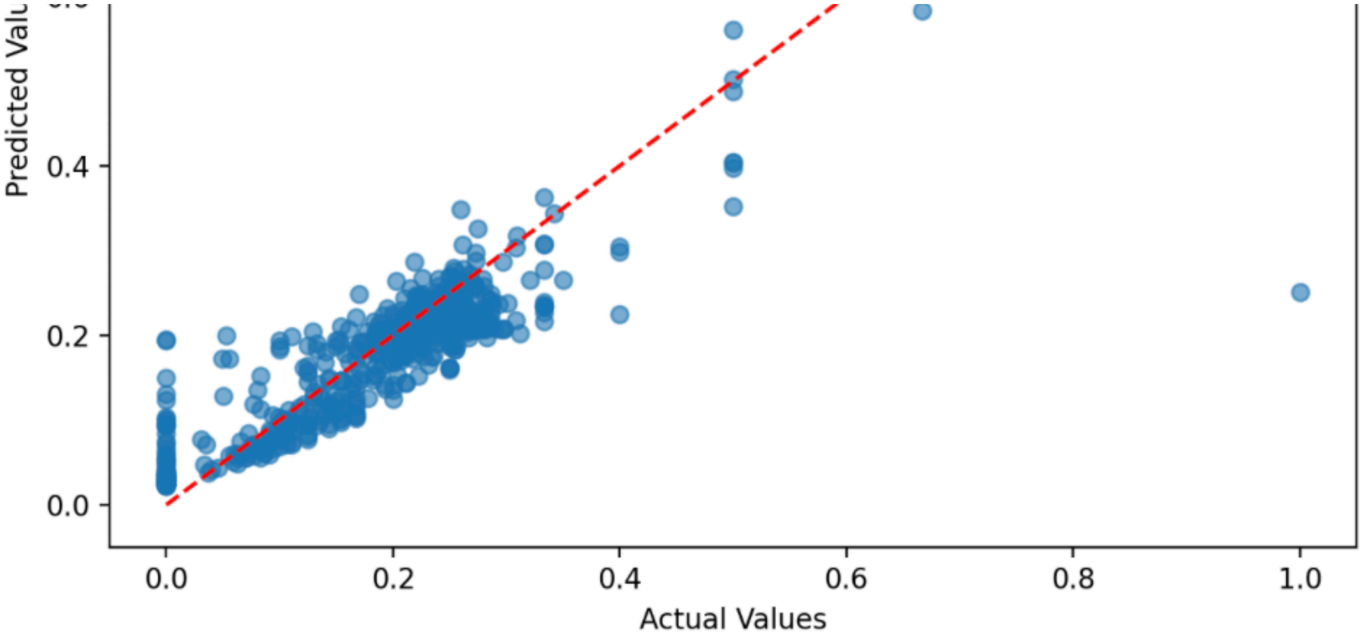


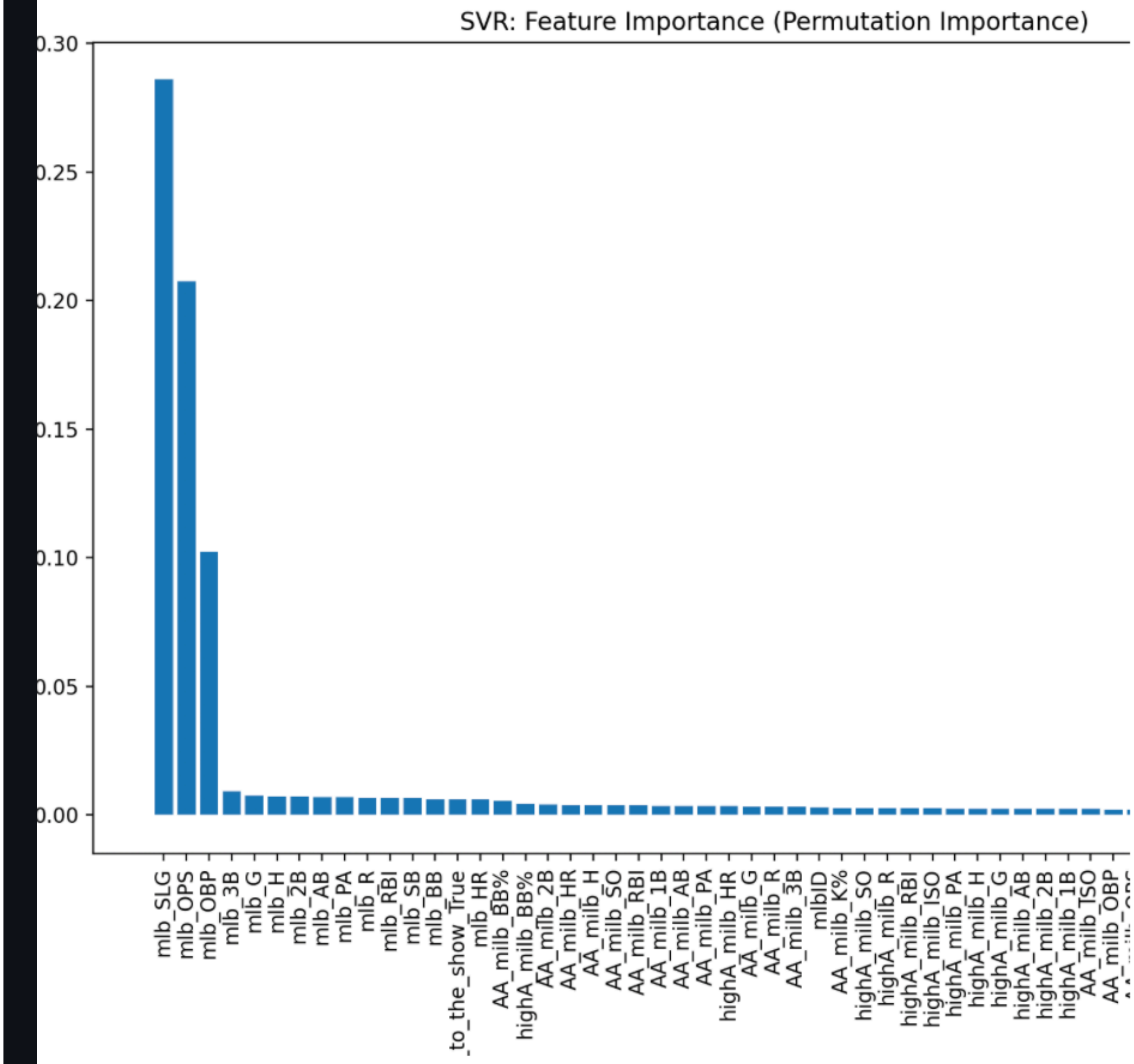
Linear Regression's metrics (MSE: 0.00190, R-square: 0.866) indicate moderate predictive accuracy: The relatively low MSE, while higher than Random Forest, still shows reasonably controlled prediction error, indicating the model maintained accuracy in predicting MLB batting averages.

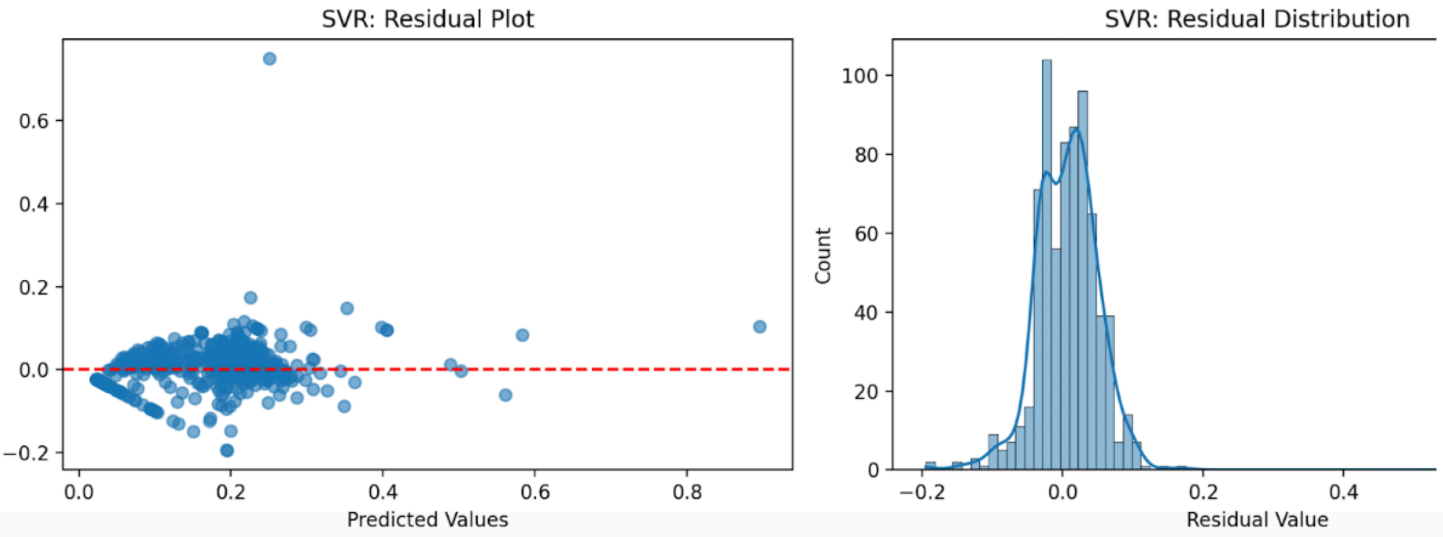
The R-squared value of 0.866 suggests the model explains 86.6% of the variance in batting outcomes. This strong performance indicates that many NCAA batting statistics have linear relationships with MLB performance. The predicted vs actual values plot shows consistent performance across different ranges, with tighter clustering in the 0.0-0.4 range where most batting averages fall. The identified key feature (strongest predictors) are ISO, OPS, and SLG, aligning with known baseball statistics [3] The Linear Regression model's effectiveness can be attributed to several factors. The residual plot shows residuals roughly symmetrically distributed around zero and most errors concentrated between -0.2 and 0.2, indicating unbiased predictions. The learning curve shows stable performance after 1000 training examples, with training and cross-validation scores converging around 0.90, suggesting the model has found an appropriate balance between bias and variance. We think Linear Regression performed well because batting statistics tend to have strong linear relationships. Players performing well in college (high batting average, high slugging percentage) will often maintain proportionally similar performance levels in MLB.

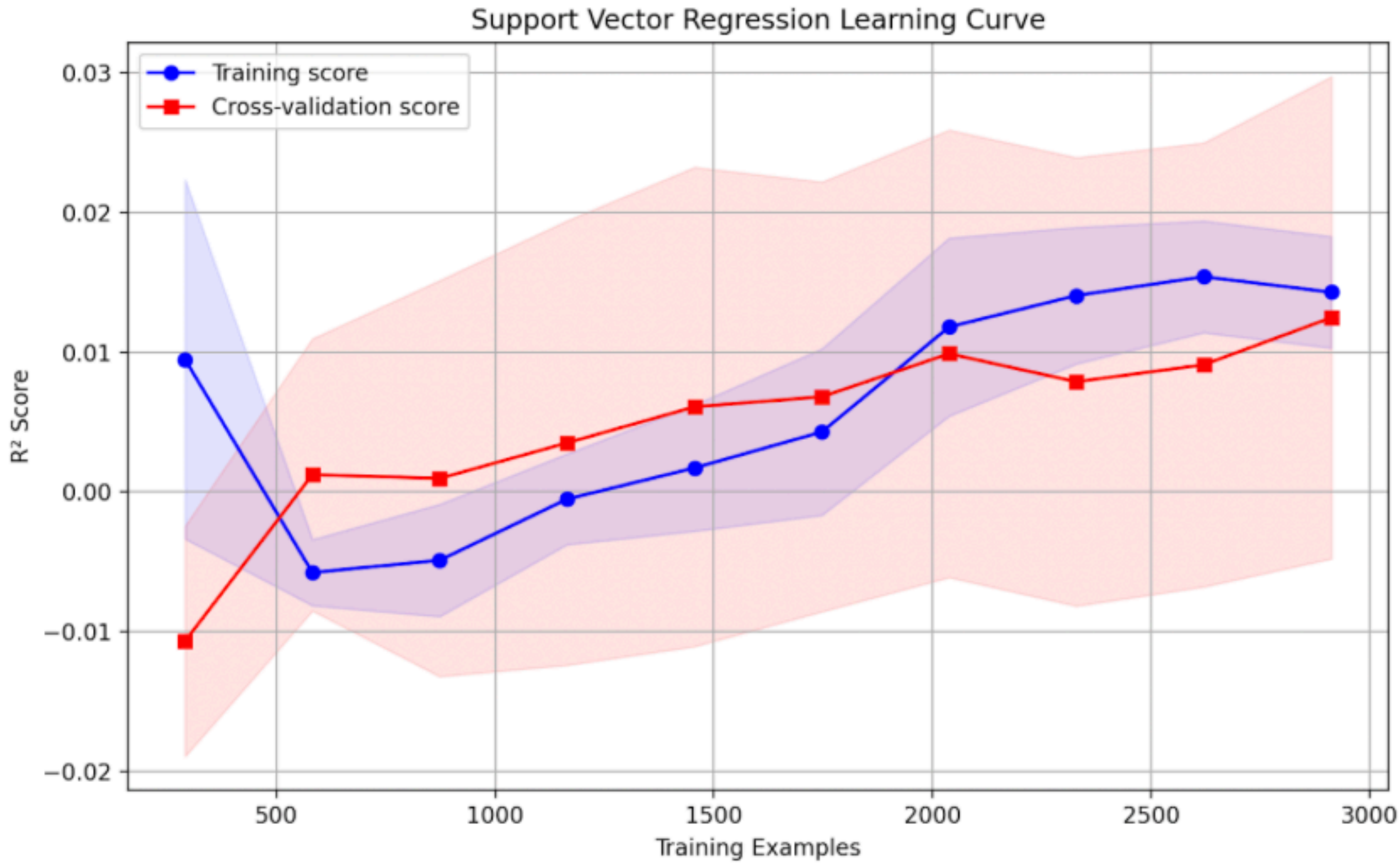
SVR




Predicted Valu
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
Actual Values


SVR: Feature Importance (Permutation Importance)
.30
.25
.20
.15
.10
.05
.00
mlb_SLG
mlb_OPS
mlb_OBP
mlb_3B
mlb_G
mlb_H
mlb_2B
mlb_AB
mlb_PA
mlb_R
mlb_RBI
mlb_SB
mlb_BB
to_the_show_True
mlb_HR
AA_milb_BB%
highA_milb_BB%
AA_milb_2B
AA_milb_HR
AA_milb_H
AA_milb_SO
AA_milb_RBI
AA_milb_1B
AA_milb_AB
AA_milb_PA
highA_milb_HR
AA_milb_G
AA_milb_R
AA_milb_3B
mlbID
AA_milb_K%
highA_milb_SO
highA_milb_R
highA_milb_RBI
highA_milb_ISO
highA_milb_PA
highA_milb_H
highA_milb_G
highA_milb_AB
highA_milb_2B
highA_milb_1B
AA_milb_TSO
AA_milb_OBP





Support Vector Regression's metrics (MSE: 0.00266, R-square: 0.812) show lower predictive accuracy compared to both Random Forest and Linear Regression: The R-squared value of 0.812 implies the model explains 81.2% of the variance in outcomes. While still pretty good, SVR captured less of the relationship between NCAA and MLB performance than the other models. Hence, the predicted vs actual values plot shows more scattered predictions. The permutation importance analysis identified three dominant features (SLG, OPS, OBP). The SVR model's lower performance can be understood through several key factors. The residual plots reveal a wider spread of errors compared to Linear Regression, though still mostly concentrated between -0.2 and 0.2. The wider spread indicates less precise predictions overall. The learning curve shows significantly slower learning and more unstable performance, with training and validation scores not converging until around 2000 examples. SVR probably performed worse because the RBF kernel, while able to capture non-linear patterns, may be unnecessarily complex for this problem where Linear Regression already showed strong linear relationships exist between college and professional statistics. The model's heavy reliance on just three features also suggests it might be underfitting the data by not effectively utilizing the full range of available statistics.

Model Comparison and Tradeoffs: Random Forest offers the highest prediction accuracy but requires more computational resources and training time due to its ensemble nature. Its "black box" nature might make it harder to explain predictions to scouts or team managers who probably would love to understand why a player is predicted to perform well. Linear Regression provides a good balance. Its main advantage is interpretability - we can directly see how each college statistic influences MLB predictions through feature coefficients. It's also efficient and trains quickly. However, it's limited to capturing linear relationships. It will miss more complex patterns in player development and, hence, have a cap on its prediction performance. SVR trades off simplicity for flexibility. While capable of capturing non-linear patterns, this flexibility comes at the cost of higher computational complexity and more difficult parameter tuning. However, it should potentially outperform Linear Regression on more complex player development patterns if properly tuned. In a practical scouting context, the choice between these models might depend on factors like available computational resource and need for interpretable predictions

Next Steps

We would want to figure out how the real-world effect of the models and the machine learning visualization we created would be like. It would be nice to look at and acommodate a few more models if we could and then figure out the comparison of those ones as well. Since we have only incorproated 3, we would like to see if any other ones would be ideal for what we are trying to do as well, in a time effective manner of course. Real-world implications are important for our work and we want to make sure that this could eventually be used in baseball of some sorts.

Click here to view the Gantt Chart

Gantt Chart for Baseball Project (Table Format)

	Task	Person	Start Date	Finish Date
0	Project Proposal	Josh	2024-09-27	2024-10-04
1	Introduction & Background	Josh	2024-09-27	2024-10-04
2	Problem Definition	Josh	2024-09-27	2024-10-04
3	Methods	Anthony	2024-09-27	2024-10-04
4	Potential Results & Discussion	Sai	2024-09-27	2024-10-04
5	Video Recording	Arnav	2024-09-27	2024-10-04
6	GitHub Page	Sai	2024-09-27	2024-10-04
7	Data Sourcing and Cleaning (Model 1)	Steven	2024-10-07	2024-10-15
8	Model Selection (Model 1)	Josh	2024-10-15	2024-10-18
9	Data Pre-Processing (Model 1)	Sai	2024-10-18	2024-10-25

Contribution Table

Project Tasks Completed

	Person	Tasks Completed
0	Josh Forden	Data Sourcing & Cleaning, Slides, Video

1	Anthony Pastrana	ML Models 2+3 (Linear Regression + Support Vector Regression), References
2	Saisaketh Koppu	Streamlit + Github and Results, Composition
3	Steven Hao	ML Model, Visualization/Metrics
4	Arnav Chintawar	Data Preprocessing, Write-up, Visualization

References

1. J. Zimmerman, “A new way to look at College Players’ Stats,” The Hardball Times, https://tht.fangraphs.com/a-new-way-to-look-at-college-players-stats/ (accessed Oct. 4, 2024).
2. “Baseball statistics, ” Baseball Reference, Feb. 23, 2024. https://www.baseball-reference.com/bullpen/Talk:Baseball_statistics
3. “Pythagorean Winning Percentage,” MLB Advanced Media, 2024. https://www.mlb.com/glossary/advanced-stats/pythagorean-winning-percentage
4. “Pythagorean Winning Percentage,” MLB Advanced Media, 2024. https://www.mlb.com/glossary/advanced-stats/pythagorean-winning-percentage