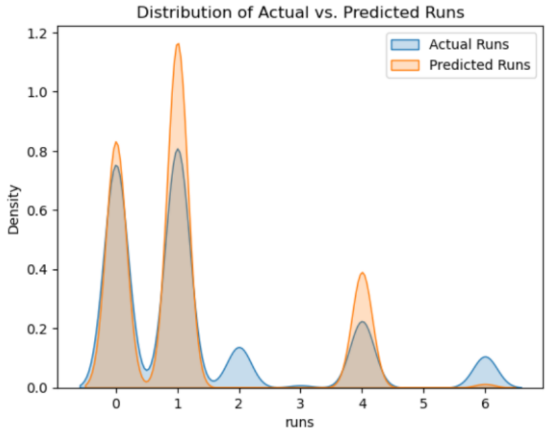
Cricviz – Take Home Assessment

1. Which Machine Learning model have you picked and why?

A. I picked an XGBoost Classifier model with added aggregate level features. When I tried a regression model, the model failed to predict boundaries (4s and 6s) completely, inspite of me trying different sets of hyperparameters and weighting methods. The XGBoost classifier was able to predict boundaries decently, and even in case of incorrect predictions, the predicted class was very close to the actual class, and shared very close probabilities. A regression model does not give us the advantage of providing us probability of classes, and hence was almost completely unusable, given its nature of predictions. I also felt coaches might understand better if I could explain the probability a delivery will be hit for a boundary or probability a delivery will be a dot ball.

2. How are you evaluating the results of your model? (e.g. What metrics and visualisations are most useful)
   A. A combination of metrics were used to evaluate my model. Along with a simple accuracy, I used log-loss and multi-class roc-auc scores, both of which indicate how well the model distinguishes between the boundaries. 0s and 1s are being captured quite well, but given the data, the model seems to struggle to identify boundary balls. However, this means that if the model classifies a ball as a boundary ball, it must be taken seriously!

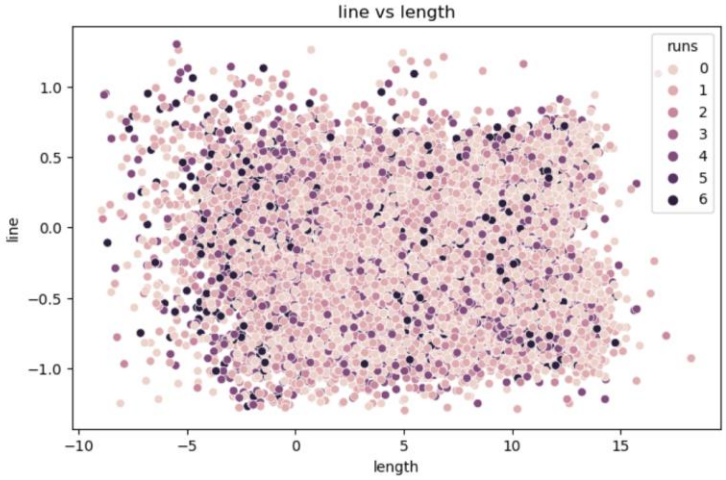
3. Does the variance of your model's output distribution match the variance of the target distribution (runs)?
   A. No, the variance of the model's output fails to match the variance of the output. The variance of the actuals is 2.65, while variance of predictions is 1.91, which means the ratio of variances is about 0.72. This indicates that the predictions are too concentrated.



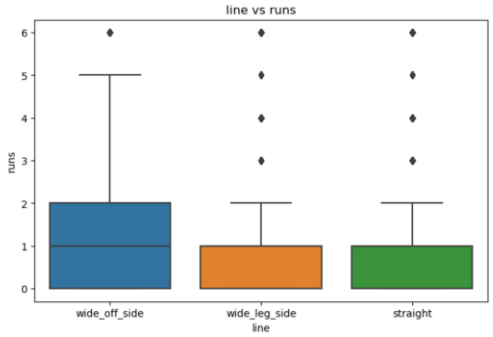
The training dataset lacks enough examples of high run values, and hence, the model struggles to predict those. However, I tried to fix this by passing a weights parameter during the model fit, which partially fixed the problem (variance ratio improved drastically). Still, it is difficult for the model to predict boundaries (especially 6s). Even after trying sample weighting, synthetic data (SMOTE), different objectives, hyperparameter tuning, and ensemble methods, the prediction variance still remains lower than the actual variance. So, it's likely inherent to the data itself.
If certain run values (like 2, 3, 5, 6) are naturally rare, even the best model cannot predict them well because might be no clear patterns distinguishing when those runs occur.

4. Let's say you're working alongside some cricket coaches in the IPL. How would you explain the learnings of your model to a coach, about the best areas to target as a bowler?
   A. Some insights can come from simple tasks such as EDA, while others from post-modelling shapely analysis.

      EDA based insights:

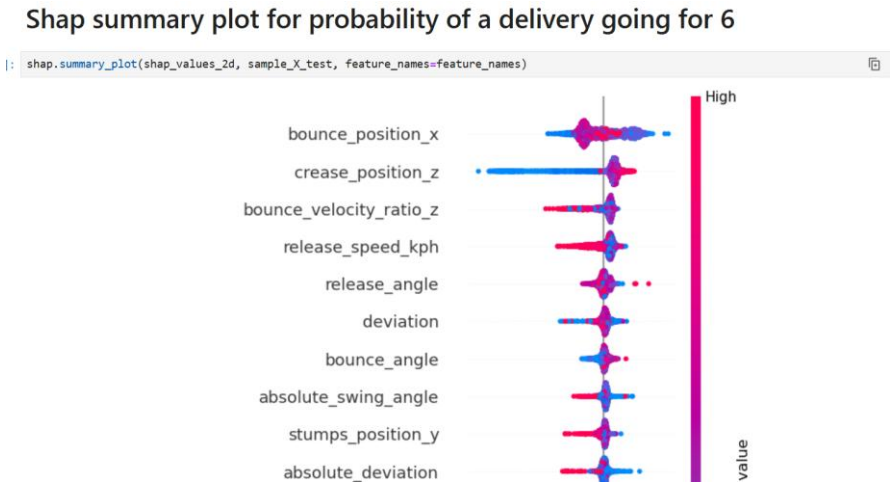


      The plot above views datapoints based on their line and length. It can be seen that 4s and 6s are mostly present towards either left or right, indicating larger absolute lengths (wider lengths) produce larger number of boundaries. **So, the suggestion to the coach would be to ensure the bowlers stick to straighter length.**



	runs		
	mean	median	std
delivery_line			
straight	1.211618	1.0	1.551581
wide_leg_side	1.346046	1.0	1.625141
wide_off_side	1.355824	1.0	1.694998

The two visuals above also support that **straighter balls have a lower mean number of runs with a lower standard deviation as well**.

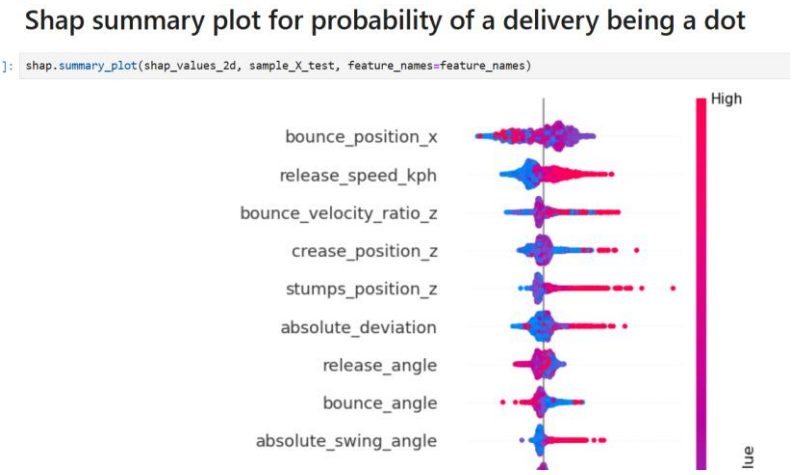


A SHAP summary plot is tells us how different features impact probabilities. Red indicates a high value for the corresponding variable, while blue indicates a lower value for the corresponding variable. Presence of a point towards the right indicates that the point pushes the probability up, and towards the left indicates that the point pushes the probability down. Above is a plot that represents the impact of variables on the probability a ball will go for 6.

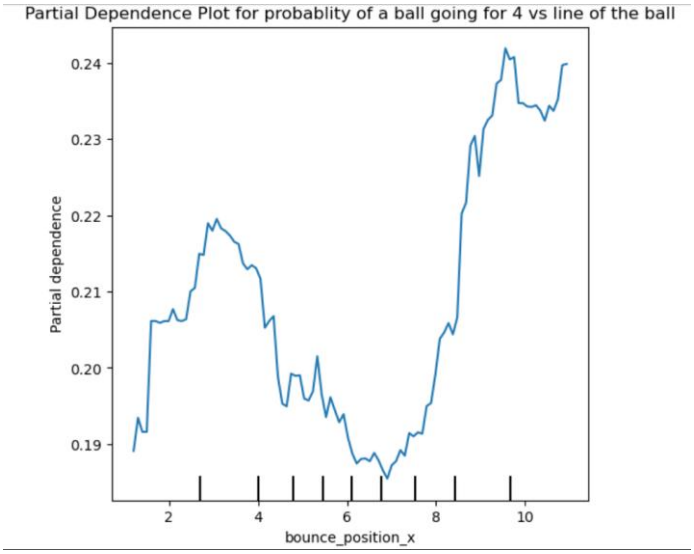
Crease position z - a lower height of the ball at crease level, significantly decreases the chances of being hit for 6. A slightly higher delivery at the crease is more likely to be hit for 6. Since a lower height at the crease level reduces the chances of the ball being hit for six, **bowlers should focus on maintaining a lower trajectory, especially in death overs or when defending a total.**

Bounce_velocity_ratio - **Hit the Pitch Harder** – A higher bounce velocity ratio means the ball retains more speed after pitching, making it harder for batters to get under and loft the ball. **Bowlers should aim to hit the pitch hard to ensure the ball skids through rather than slowing down.** On surfaces

where the ball slows down significantly after pitching, batters find it easier to time big shots. Bowlers should adjust their strategies accordingly—**either by bowling fuller to prevent batters from setting up for big shots or by using cutters and slower balls to exploit inconsistent bounce.**



The above shap plot exhibits impacts of variables on the probability of a ball being a dot. It can be seen that higher release speeds, bounce velocity ratios, absolute deviation and swing angles are associated with higher probability of a ball being a dot. **Bowling fast, especially when the pitch supports seamers and spinners is highly recommended!**



The partial dependency plot above also shows that balls bowled at 6-7 meter mark have the lowest probability of being hit for a boundary. So bowling in that 6-7 meter mark is the way to go!

5. What future steps might you want to explore if you had another week of coding for this model?

- Feature Engineering wise, I tried everything I wanted to. However, I would probably try other sampling techniques to balance the data such as ADASYN, Borderline-SMOTE, Tomek Links etc.
- I would have tried ensembling techniques – different models to predict different classes – one model to predict 0,1 and 2, and another to predict 4 and 6. Then I would combine the two models
- Predicting expected runs for a given delivery is extremely challenging when only ball-related metrics are given. Some context around the batter, situation of the game, field placement etc. would significantly improve results. Even one additional feature that would give some more context would have been helpful. So, given more time, I would love to include such variables in my modelling and analysis.
- While I fine-tuned the model to some extent, with more time, I would experiment with Bayesian Optimization, Hyperband, or genetic algorithms to further optimize hyperparameters for better predictive performance.
- I would conduct more advanced post-model analysis by using 2D partial dependence plots and SHAPELY dependence plots, that take longer and more computational capacity.