Expected Pass Model

Carter Bouley

The Data

Initial Variables:

- Team, Game, Player, Home, Away IDs
- Home & Away Full Time Score
- Half, Minute, Second
- Outcome of Pass
- Start x & y, End x & y of pass
- Header, Cross, Corner, Thow, Goal Kick (or throw), Free Kick

The Data

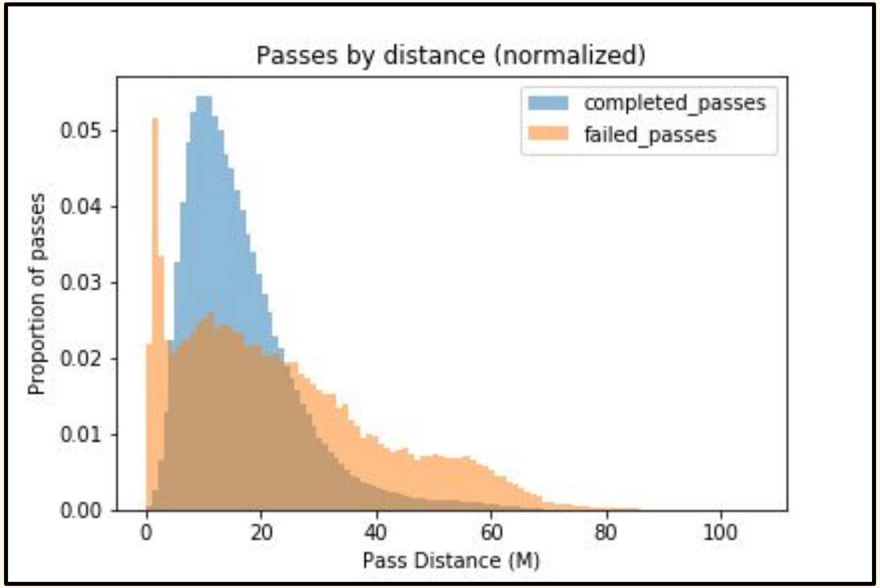
Teams	20
Games	380
Players	716
Pass Attempts	358,783
Completed Passes	284,057
Incomplete Passses	74,726

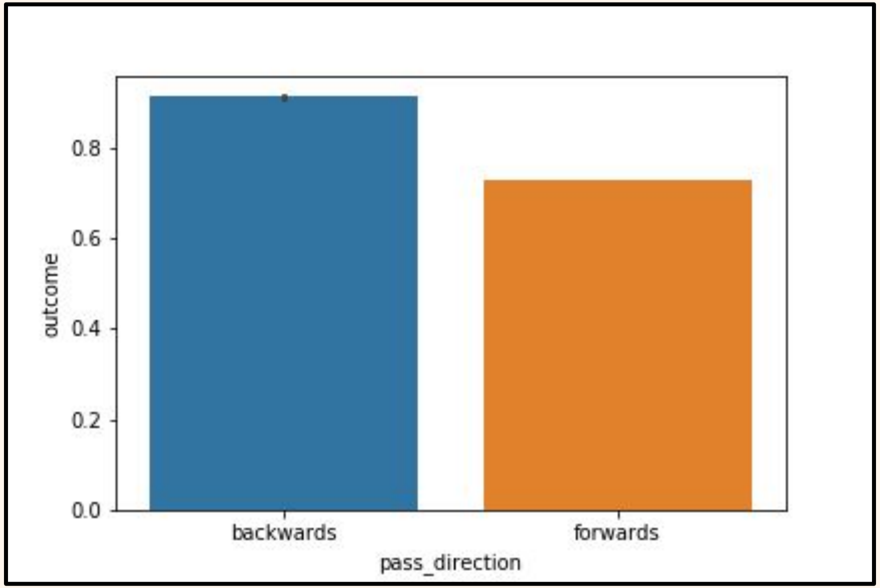
Pass Success Rate: **79.17%**

Initial Feature Additions

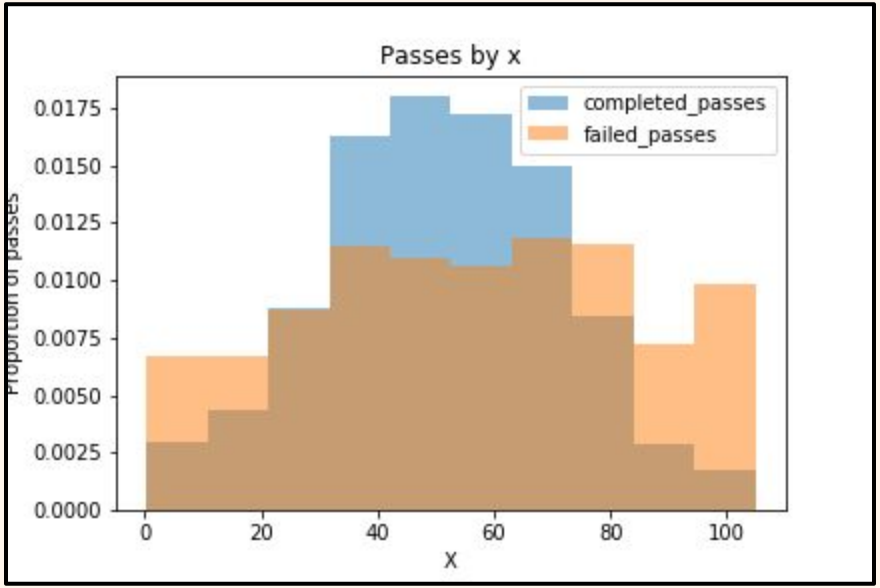
- Pythagoras' Theorem to calculate pass distance.
- Home & away location metric for each game.
- Forwards or backwards for pass.
- Game result (home, draw, away)
- For the team currently in possession, their game result. (win, draw, loss)

Initial EDA

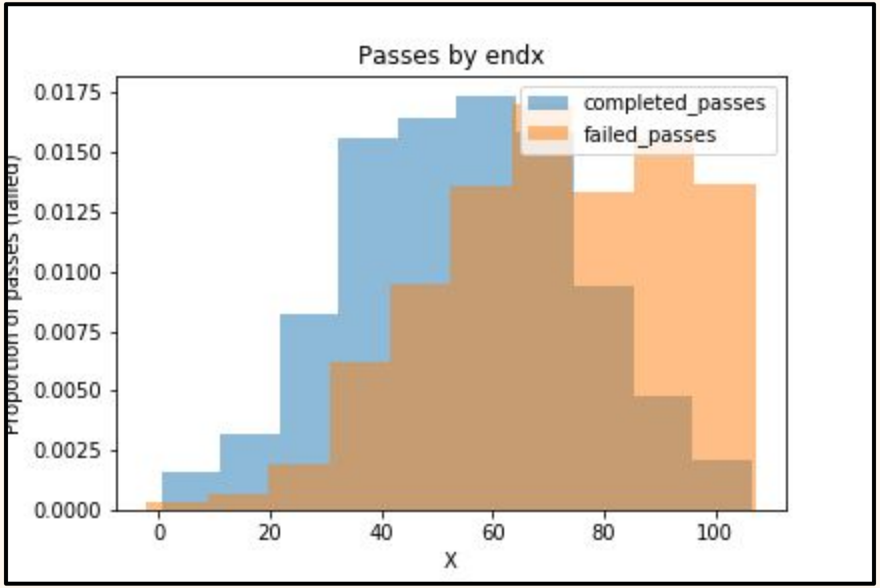

Passes by distance (normalized)
completed_passes
failed_passes
0.05
0.04
0.03
0.02
0.01
0.00
Proportion of passes
0
20
40
60
80
100
Pass Distance (M)


0.8
0.6
0.4
0.2
0.0
outcome
backwards
forwards
pass_direction

Initial EDA

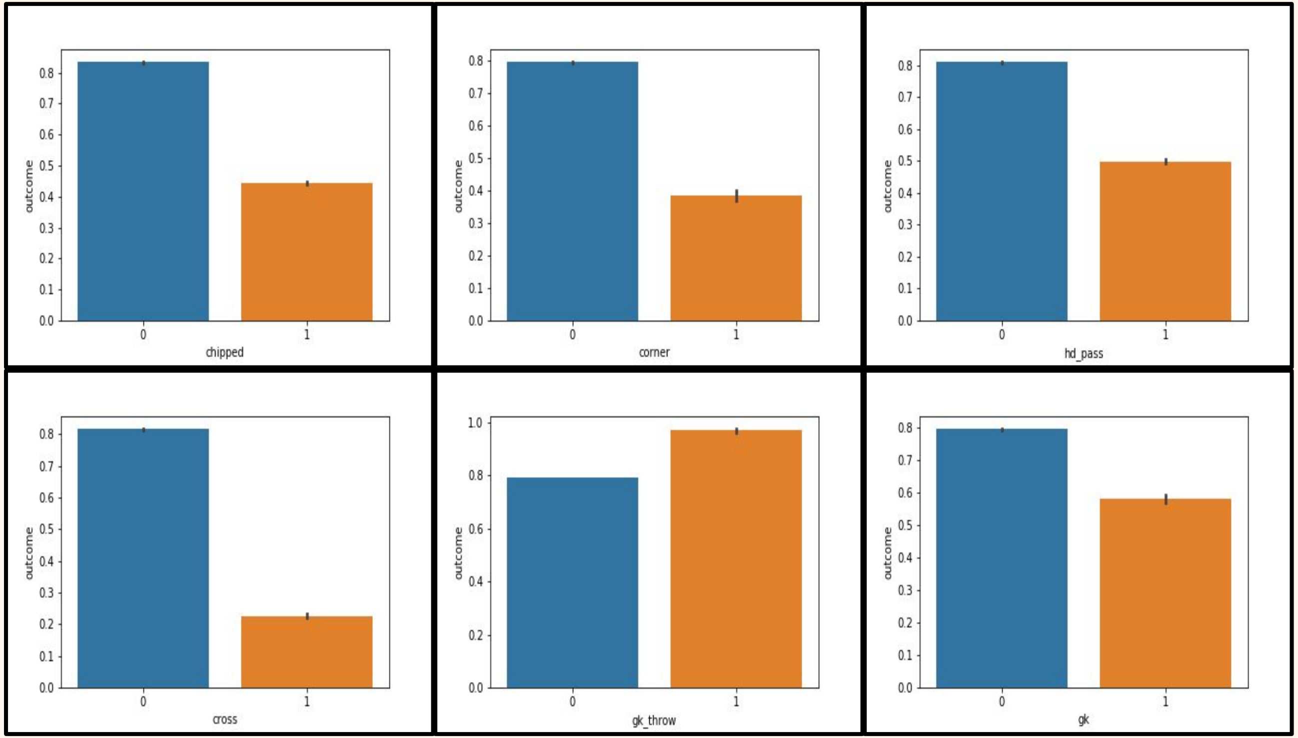


Passes by x
completed_passes
failed_passes
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
20
40
60
80
100
X



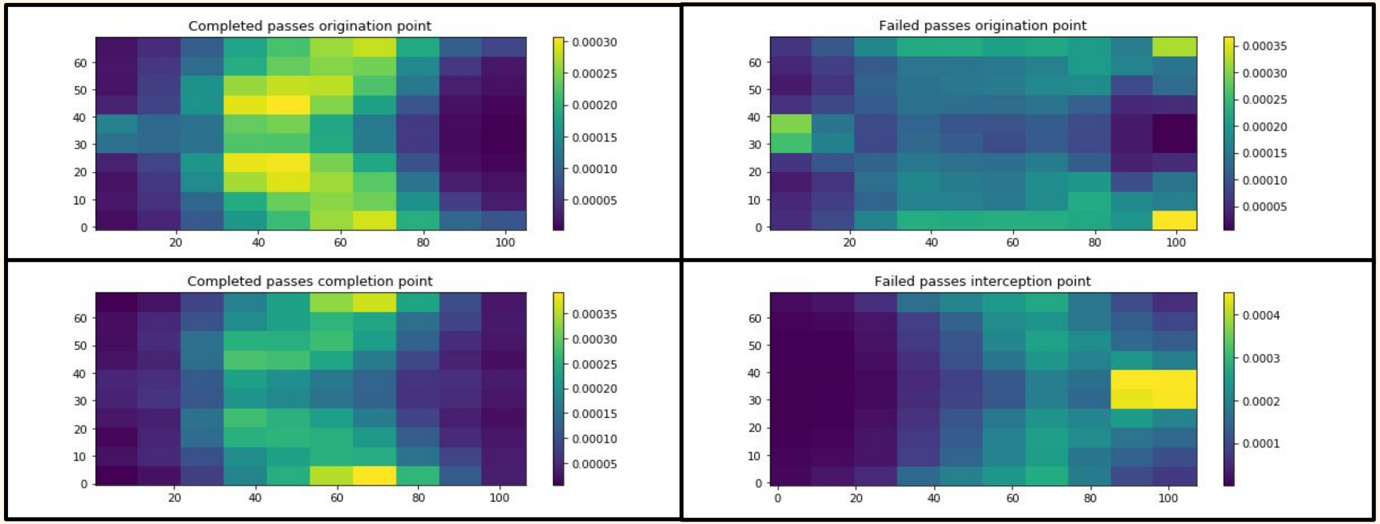
Passes by endx
completed_passes
failed_passes
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
20
40
60
80
100
X

Initial EDA



Special feature passes generally contain a lower proportion of completed passes, ***unless*** they are goalkeeper throws.

Initial EDA



- Completed passes generally occur around midfield.
- Failed passes generally start either near a teams own box, or near opponents corner.
- Failed passes are generally stopped near the opponents own box.

Initial EDA

Attacking Team Direction of Play

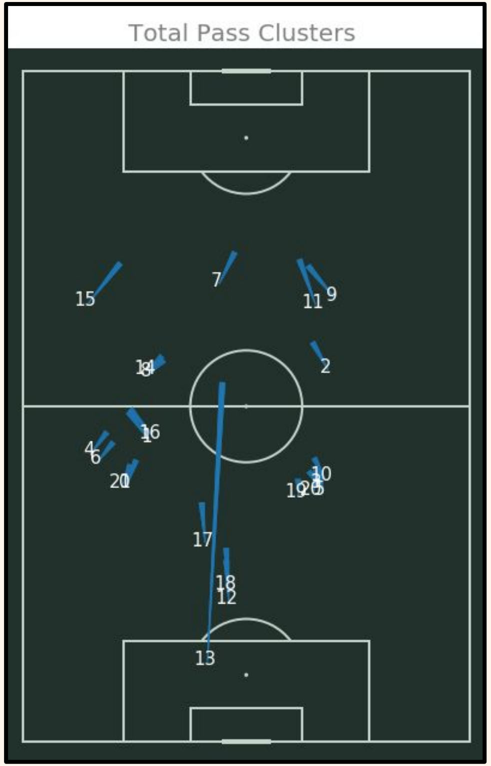
Addition of grid labels as a dummy variable to get a secondary look (other than simply x and y coordinates) into areas where a team performs particularly well or poorly.

Unsupervised Learning

- ***K- means*** clustering is simple to implement.
- It is relatively fast when compared to hierarchical methods.
- Algorithm scales to large datasets.
- The algorithm adapts to new examples reasonably easily.

- **Choosing K:**
  - Elbow method imprecise and plot revealed little
  - Optimal K based on silhouette score was 4 however issue when visualizing clusters.
  - Settled on 22 = accepting the reduced silhouette score as a trade off for increased interpretability.

Unsupervised Learning



Total Pass Cluster Analysis:

- 22 cluster centroids, including starting and ending x and y position
- Gives a rough idea of the different kinds of passes that exist, along with how they look on a pitch.
- Lots of clustering around mid-field - suggesting much of the passing takes place there.
- This aligns when assessing our origin points, where **55%** of passes across the whole season originated in points **4**, **5** and **6**, despite only taking up a third of the pitch.

Supervised Learning

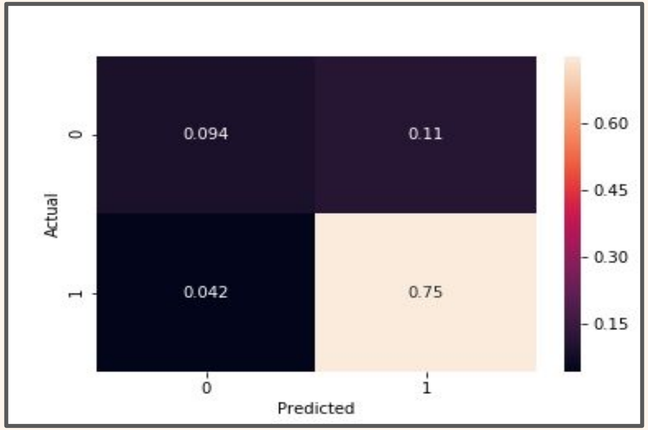
Baseline (Dummy Model):

- Selects majority class every time
- In the test set, 71,001 passes were completed, and 18,695 failed.
- Selecting pass complete every times gives an accuracy of **79.3%**
- However, leads to a high log loss (**7**) and AUC of **0.5,** rendering it worthless.

Supervised Learning

Logistic Regression:

- Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
- Easy to implement, interpret and very efficient to train.
- Main limitation of Logistic Regression is the assumption of linearity between the log odds dependent variable and the independent variables.



- Accuracy Score : **0.84**
- Precision Score : **0.87**
- Recall Score : **0.95**
- F1 Score : **0.91**
- Log loss :**0.36**
- AUC**: 0.86**

Supervised Learning

Logistic Regression Insights:

- With this data, x had a positive coefficient, suggesting that as x increases the probability of a completed pass increases.
- However, end_x had a large negative coefficient, which suggests that the end location of a pass being further up the pitch reduced the probability of completion
- Crosses, Headers, Corners and Throws all had negative coefficients
- Based on the nine-level split of the pitch,end sections had larger coefficients than origin sections, (and were all significant at 95%) suggesting pass end location matters more than beginning of a pass.

Pass Difficulty:

- Largest pass end coefficients were section **2**, and interestingly section **9**. The lowest was section **8.**

This is in comparison to the dropped section **5**.

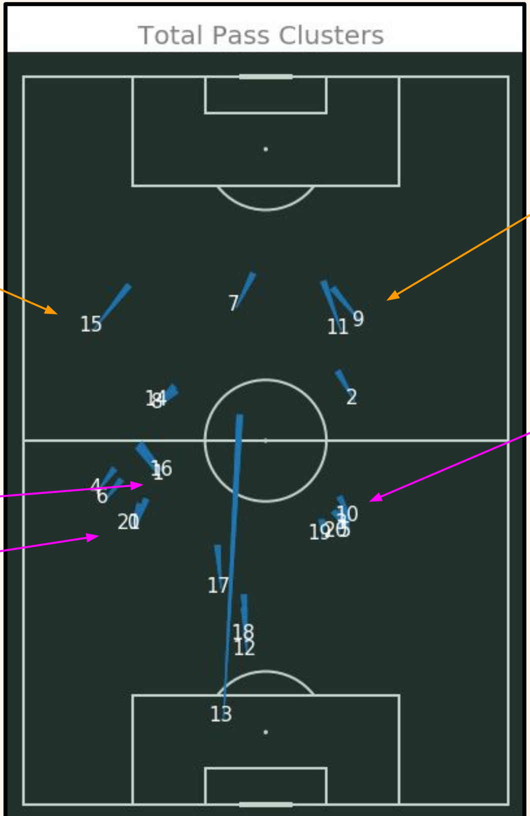
Supervised Learning

Cluster Insights:

- Clusters which had the highest positive coefficients were cluster **1**, **20** and **5**
- Clusters with the lowest negative coefficients were **15**, and **9**
- Not all clusters were significant predictors.

Model Iterations:

- Upsampling failed passes to deal with class imbalance
    - Increased precision at cost of recall
- Hyper-parameter tuning
    - Penalty & C tuned
    - No performance increase



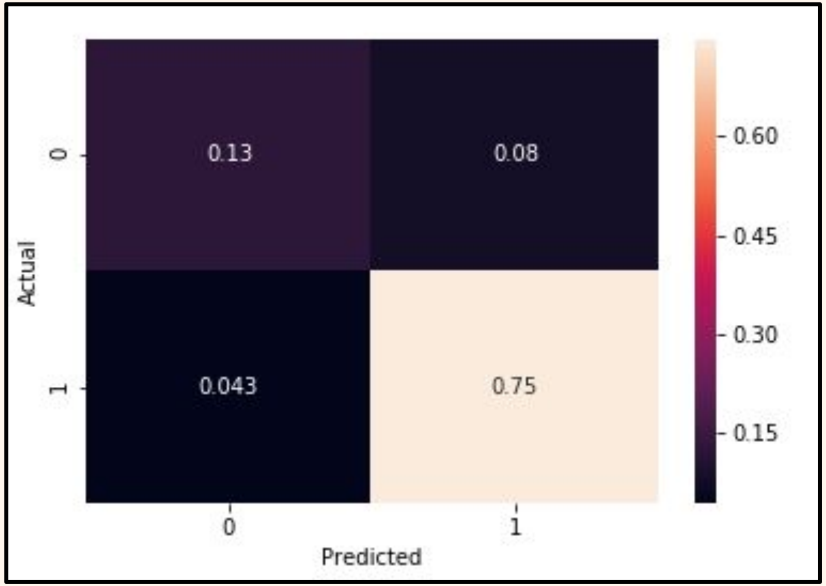
Supervised Learning

XGBoost:

- Ran with Hyper-parameter Tuning using Randomised search with cross validation

- No coefficients to use for coaching, however much better accuracy with less loss

- Accuracy Score : **0.88**
- Precision Score : **0.9**
- Recall Score : **0.95**
- F1 Score : **0.92**
- Log loss :**0.28**
- AUC**: 0.92**



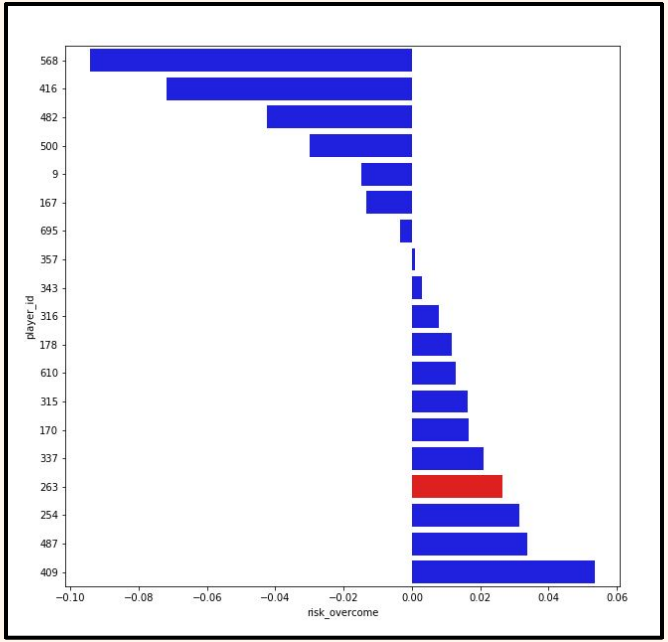
Player Analysis

Pass Risk:

- Use XGBoost to predict the probability of a pass being made.
- If the probability is 0.2, and a pass is complete (1) I add 0.8 to the risk overcome metric
- This enables the ability to compare players across the risk taken, and overcome in passing

Player Similarity:

- Store every player which shares the same largest cluster



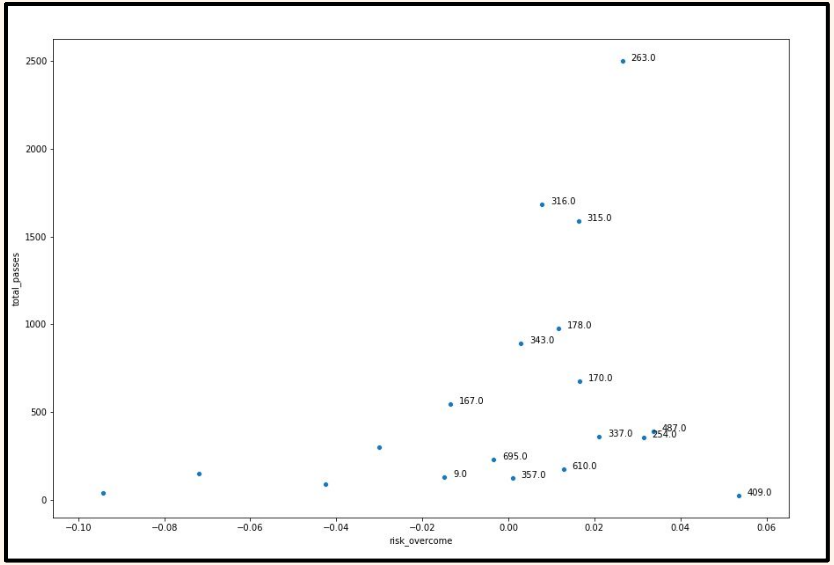
Player Analysis

Model Insights:

- Player 263's ability to overcome risk in passing is slightly above average when compared to similar players

- However, he attempts many more passes than similar players over the season.

- Other labelled players could be worthY replacements. Player 409 is potentially underplayed/undervalued.

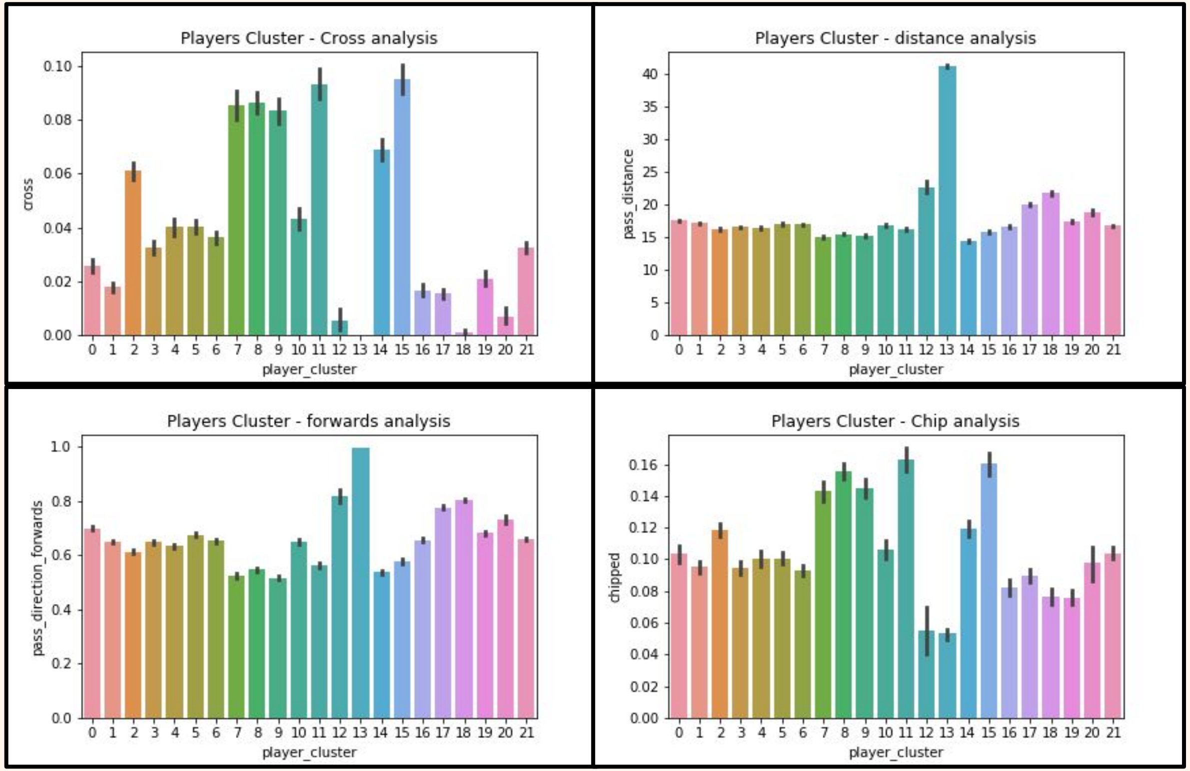


Player Analysis

Passing Styles:

- Through selecting a players most common cluster, we can group similar players

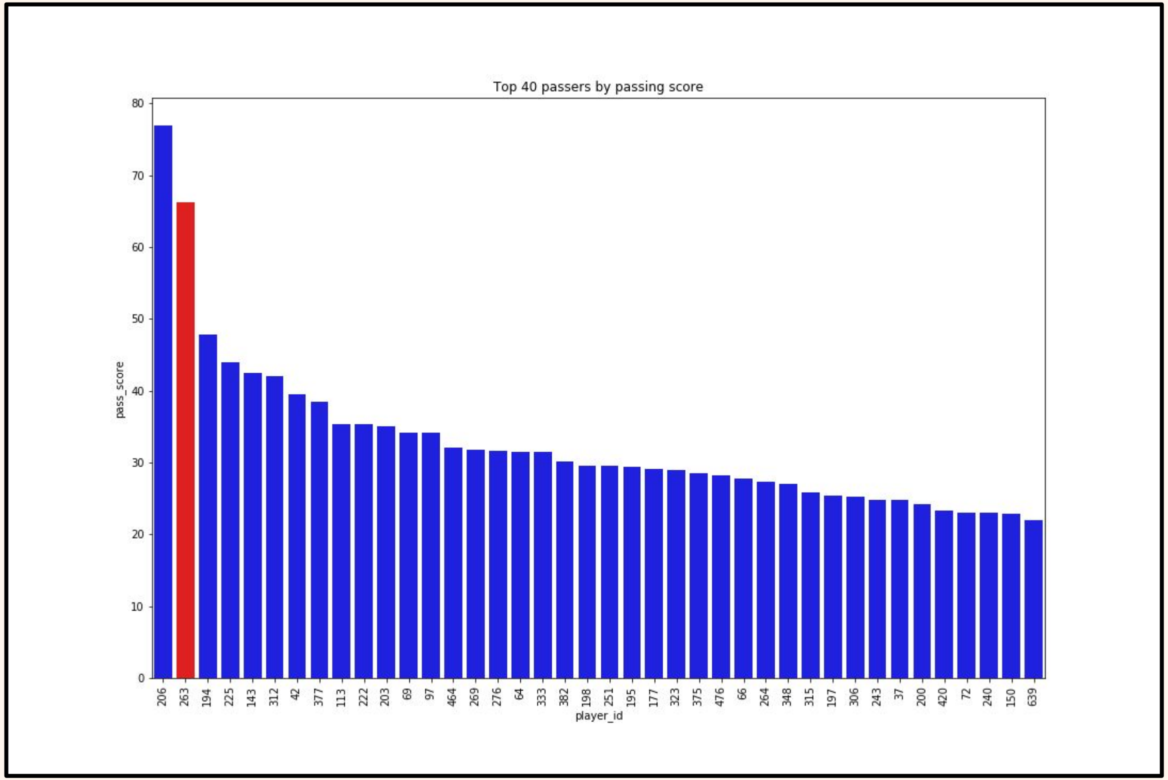
- Through doing this we can see the different styles that these players use to pass.



Top Players

Passing Score:

- Both overcoming risk, and number of passes are important metrics.
- Multiplying these together generates a passing score.
- Our initial player 263, lies within the top 40 passers of the season.



Further Work

Limitations:

- Ranking top passers like this assumes risk is good - No concept of value.
- Clustering player types based solely on their most popular pass choice - over-simplification of the game.

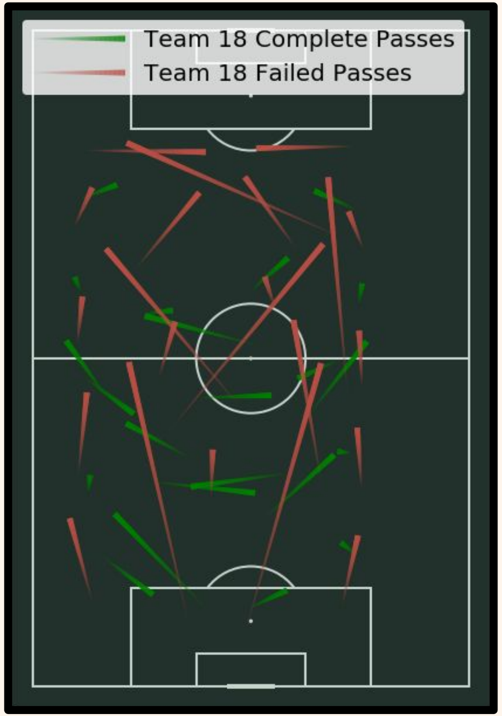
Improvements:

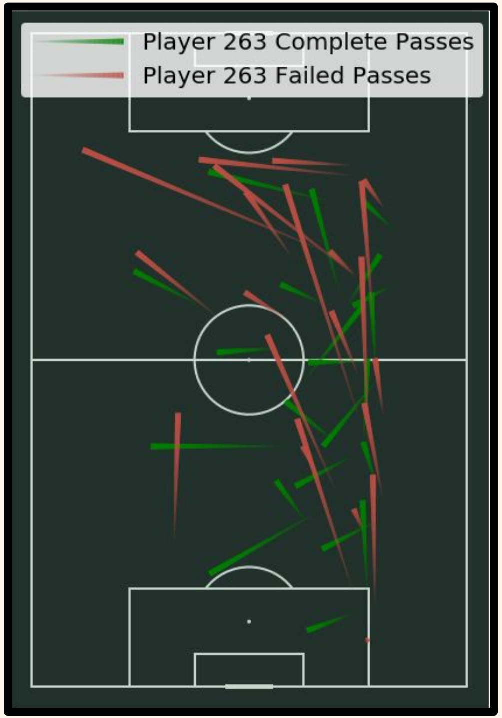
- Building a concept of value, or reward per pass will help define ‘good’ passing ability
- Building similarity scores between players rather than this simple methodology
- Building more granular models at the team level, rather than the league.

Further Team Analysis

Individual Team & Player Clustering

- Teams have varying styles of play
- One opportunity could be perform this clustering at the team level (instead of the league)
- This could lead to a deeper understanding of either your own, or opponents passing styles.
- Building pass risk models at a more granular level may be more applicable.





Questions