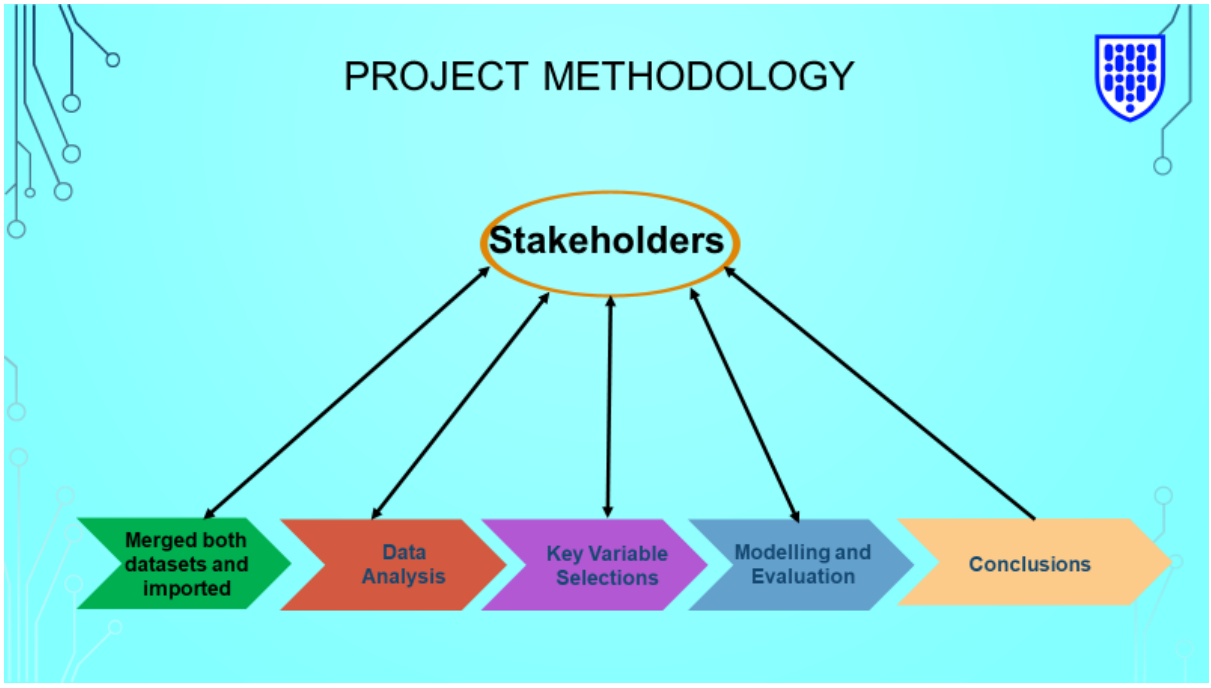
Capstone Project Document

DSIA – Data Science and AI Course
Part-time – 2020-12-01

Rohit Subramaniam

Process Overview



Problem Statement

- **What is the problem or the opportunity that the project is investigating?**

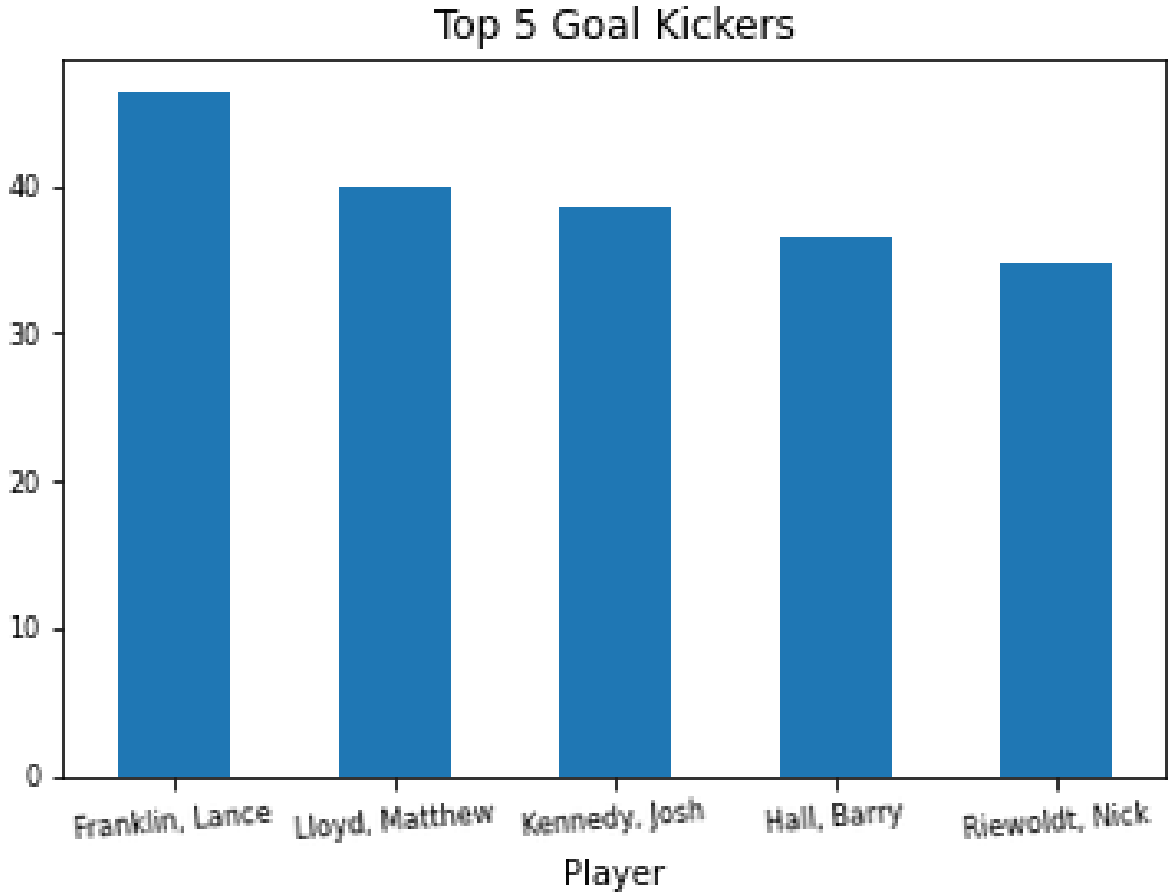
To understand the impact of player position and other performance indicators such as number of kicks on goal kicking performance.

- **Why is this problem valuable to address?**

Over the years, the game has become very fast paced and all Australian football players are expected to play in multiple positions and kick goals irrespective of their primary role.

- **What is the current state (e.g. unsatisfied customers, lost revenue)?**

AFL (Australian Football league) is a billion-dollar industry full of elite talented players.



Top 5 goal kickers in the last 20 years (1999 – 2019)

- **What is the desired state?**

  Players who can kick more than 2 goals are deemed to be goal kickers given how the game has changed. This is because players from any position can kick 2 goals in today's game but not every player can kick more than 2 goals in a game.

- **Has this problem been addressed by other research projects? What were the outcomes?**

  The review of sports related research has not addressed this problem specifically. Hence, the problem was deemed appropriate to examine.

Industry/domain

- **What is the industry/ domain?**

  Sport – Australian Football League

- **What is the current state of this industry?**

  AFL is a billion-dollar sports industry that is expected to keep flourishing in the future. Innovations will continue to take place to look at players' comfort levels and performance.

- **What is the overall industry value-chain?**

  Australian football league coaching staff and players.

- **What are the key concepts in the industry?**

  The Australian Football League is an iconic sport of Australia and the spirit of an Australian. One of the key objectives in the AFL industry is to enhance the game from a social and cultural point of view in Australia. The game enhances the cultural diversity in the country. Players from different country of origin are brought together in a common platform to demonstrate their skills. Teams to deliver quality performance to spark entertainment for the crowd.

- **Is the project relevant to other industries?**

  Yes, to the sports and allied health industries.

Stakeholders

- **Who are the stakeholders? (be as specific as possible)**

  Australian Football League, AFL clubs and coaches, Sports Analyst and Champion Data.

- **Why do they care about this problem?**

  Kicking goals is the ultimate motive for every team. Goals determine the scoreboard which in turn determines the result of every game. But we need to determine factors that drive goal kicking performance such as the position of the players. Is it only the forward players who have stronger impact on goal kicking or can players from other positions score more goals?

- **What are the stakeholders' expectations?**

  Stakeholders would hope to see what positions make a difference in goal kicking performance. For example, utility, half back, half forward, etc.

Business Question

- **What is the main business question that needs to be answered?**

  Does player position influence goal kicking?

- **What is the business value of answering this question? (quantify value and make necessary assumptions)**

  As mentioned before, any player from the dataset who kicks more than 2 goals is deemed a goal kicker otherwise they are not. This is because in the game today, a player from any position can kick 2 goals but not every player has kicked more than 2 goals.
  When merging both datasets, any player's position that could not be found from the player position dataset was assumed to be a utility player as those players can play anywhere.

- **What is the required accuracy? What are the implications of false positives or false negatives?**

  As I am applying regression and classification, a required accuracy of more than 80% would illustrate the reliability of the model.

Data Questions

- **Which statistical models accurately reflect the factors that impact goal kicking performance?**
  All models do but more so the logistic and bagging methods as they have the highest precision, recall, accuracy, and AUC (area under curve) values.

Data

- **Where was the data sourced?**

  1.) AFL Average Game Stats for Players 1999-2019, accessed 7th September 2020 at 7:30pm, <https://www.kaggle.com/xanthangum/afl-average-game-stats-for-players-19992019/notebooks> (AFL player statistics dataset)

  2.) AFL Player Positions 2019, accessed 7th September 2020 at 8pm,

<https://www.kaggle.com/mschlitzer/afl-player-positions-2019?select=AFL_player_positions.csv> (AFL player position dataset)

- **What is the volume and attributes of the data?**

  AFL player statistics dataset has 12,742 entries and AFL player position dataset has 840 data entries (1999 - 2019).
  AFL player position dataset has only categorical variables whereas player statistics dataset has both categorical and numerical variables.
  As mentioned before, the player positions was merged into the player statistics dataset after which a target variable, “goal kicker” was created based on the assumption that if a player kicks more than 2 goals, they are a goal kicker. Another binary variable that was created was “forward player,” if the position contains the word “forward.”

- **How reliable is the data?**

  Date was taken from the Kaggle database which is high on reliability. Sports analysts draw relevant data from this database.

- **What is the quality of the raw data?**

  Very good – the data set covers all the key statistics for every single player in the Australian Football League from 1999 to 2019.

- **How was this data generated?**

  Data was generated by AFL and it was obtained from the above weblinks.

- **Is this data available on an ongoing basis?**

  The current data is updated till 2019 and it is assumed that the data will be updated on a regular basis. Generally, the dataset gets updated after every season.

Data Science Process

Data Analysis

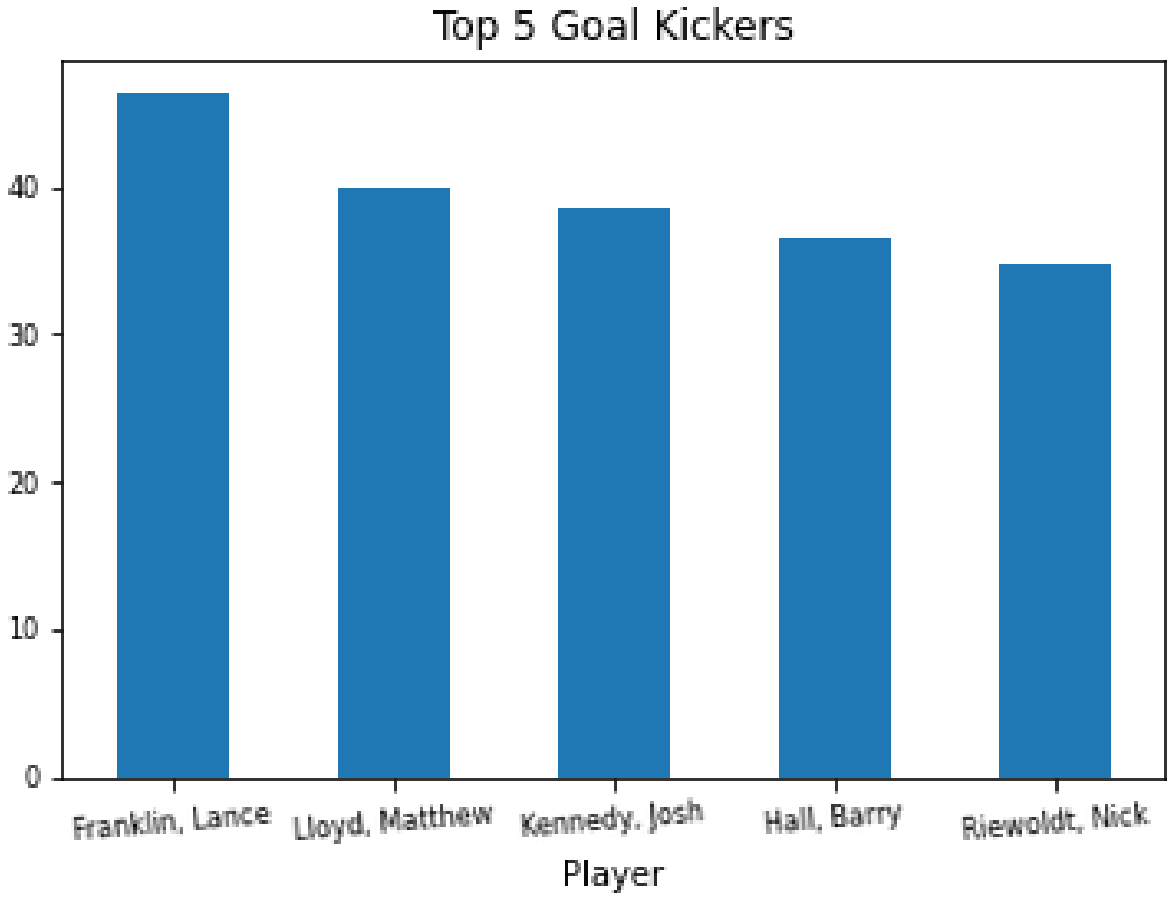
- **What data pipeline was to wrangle the raw data?**

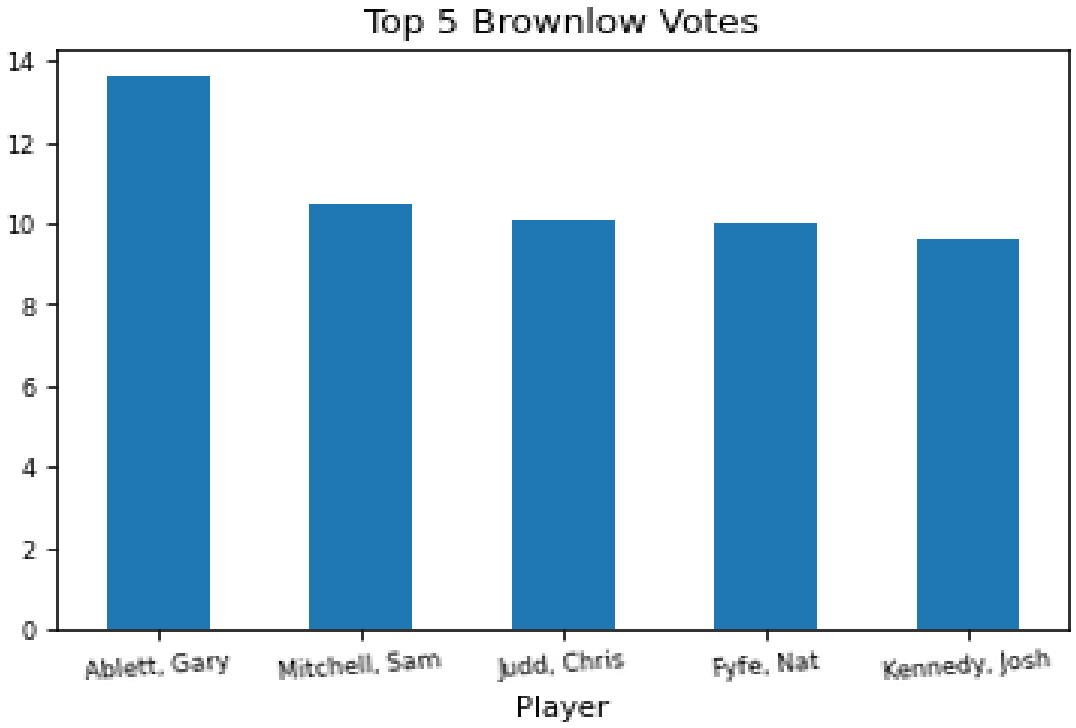
  1.) Scatterplots between variables to identify trends and correlations.
  2.) Identifying the size/shape of dataset.
  3.) Histograms
  4.) Correlation matrix

5.) Bar graphs of top 5 goal kickers, top 5 players with most disposals and so on.
6.) Linear regression plots between the variables that has the strongest positive and weakest negative correlations for the raw, training and test data.

- **What are the highlights of the Exploratory Data Analysis (EDA)?**

  Top 5 goal kickers are not necessarily the ones who are in contention to get Brownlow votes. Brownlow votes are awarded to best players by the umpires at the end of each season on the ground during each game. As we see in the graphs below, top 5 goal kickers and top 5 players for Brownlow votes over the last 20 years are different. This shows that there are other statistics such as disposals, contested possessions that also determine player's performance to win Brownlow medal. Average Marks Inside 50 and Average Goals per season for each player have the strongest positive correlation (r = 0.868). On the other hand, Average Rebound 50 metres and Average Goals per season for each player has the weakest negative correlation (r = -0.383).





Had to impute variables, disposals, and goal assists as they contained null values.

- **Is the pipeline reusable? (for example, to process future data?)**

  Yes – it should be reusable. Especially if there is a similar business problem such as does disposals influence Brownlow votes?

- **What are the intermediary data structures used (if any)?**

  Not Applicable.

Modelling

- **What are the main features used?**
  Used logistic regression, support vector machine, Naïve Bayesian classifier, Decision tree, Random Forest, Adaptive Boosting and Bagging with 6 variables included – marks inside 50, goal assists, goal kicker, contested marks, forward player and disposals.

- **Is there a subset of features that would get a significant portion of your final performance? Which features?**

  Yes – goal assists, marks inside 50s, contested marks and disposals.

- **How did you select features?**

  Features were selected based on the performance of each model in terms of accuracy score, precision, recall and Area under curve (AUC) scores. If necessary, more variables were included to improve accuracy and reliability of each model.

- **What are the models used?**

  Data was analysed using logistic regression, support vector machine, Naïve Bayesian classifier, Decision tree, Random Forest, Adaptive Boosting and Bagging with 6 variables included – marks inside 50, goal assists, goal kicker, forward player, contested marks and disposals.
  Used linear regression model to also identify variables that have the strongest positive correlations and weakest negative correlations.

- **How long does it take to train your model?**

  2-4 minutes given how large the dataset is.

- **What are the tools used? (cloud platform, for example)**

  Restricted to Excel and Python.

- **What are the model performance metrics?**

  AUC (area under curve) value, accuracy, precision, and recall.

- **Which model was selected?**

  All models were accurate but logistic and bagging classifications were selected since these methods have the highest accuracy, precision, and recall values.

Outcomes

- **What are the main findings and conclusions of the data science process?**

  In general, all models have shown that player position along with contested marks, marks inside 50, disposals and goal assists impact goal kicking performance. This will give AFL team coaching staff and players an insight what areas to focus on to be a regular goal kicker in the competition.

Implementation

- **What are the considerations for implementing the model in production?**

  Impact of player position and other key statistics on goal kicking performance can be modelled in future AFL player datasets.

Data Answer

- **Was the data question answered satisfactorily?**

  Yes –many regression models were tested that drew the same conclusion.

- **What is the confidence level in the data answer?**

  N/A as I never used confidence interval to assess model accuracy and reliability.

Business Answer

- **Was the business question answered satisfactorily?**

  Yes. Player position does have an impact on goal kicking performance along with other factors such as goal assist, contested marks, marks inside 50 and disposals.

- **What is the confidence level in the business answer?**

  N/A – did not need to use confidence interval to answer my business question.

Response to stakeholders

- **What are the overall message and recommendations to the stakeholders?**

  The positions that players play have a huge impact on goal kicking performance. These results should help all stakeholders, especially the AFL coaches and staff of each team to implement different game plan strategies to ensure that the main goal kickers achieve the most from the position they are playing in. For AFL club coaches and staff whose team have been weak with goal kicking, disposals, marks inside 50, goal assists and contested possessions should recruit players in the off season who are highly skilled in these areas.

End-to-End solution

- **What is the overall end-to-end solution to use the model developed in the project?**

  In addition to position marks inside 50 metres, goal assists, disposals, goal assists and contested marks play a big role in goal kicking performance.

References

1.) AFL Average Game Stats for Players 1999-2019, accessed 7th September 2020 at 7:30pm,
<https://www.kaggle.com/xanthangum/afl-average-game-stats-for-players-19992019/notebooks> (AFL player statistics dataset)

2.) AFL Player Positions 2019, accessed 7th September 2020 at 8pm,
<https://www.kaggle.com/mschlitzer/afl-player-positions-2019?select=AFL_player_positions.csv> (AFL player position dataset)