Introduction

This report focuses on studying the Sydney Swan's fans. A survey has been conducted to collect data, and will be used to generate recommendations for improving Sydney Swan's business performance. The data has all been cleaned by using the .isna() function to detect missing values and the .dropna() function to drop these missing values.

The following table demonstrates the questions, variables, and machine learning techniques used in this report, and would be discussed in sequence.

Questions	Target Variable(s)	Response Variable(s)	Model Name(s)
Who are the current loyal fans of Sydney Swan? What will their future support for the team be?	Q14	Q21 & Q22	Random Forest
Who are the FWFs? What factors make them loyal to the Sydney Swan?	Q14	Q6 - Q9	Logit Model
What will affect overall members' future loyalty?	Q15	Q15.1 - Q15.10	Random Forest, SVM
Are those who attend games with families willing to pay 10% more in membership fees?	Q2.1 - Q2.10	Q15.5	XGBoost Model

Analysis 1

Questions	Target Variable(s)	Response Variable(s)	Model Name(s)
Who are the current loyal fans of Sydney Swan? What will their future support for the team be?	Q14	Q21 & Q22	Random Forest

McKinsey & Company (2017) identified their customers who ranked in the top quartile as 'loyal', and discovered that the top-performing loyalty programs indeed had a positive impact on the customers. Hence, we would like to analyze every respondent's responses on question 14, and set the top quartile as a benchmark to classify the current loyal fans of Sydney Swan.

Question 14 contains six sub-questions, each using a 7-point likert scale in which a rating of 7 indicates strong agreement, while a rating of 1 indicates strong disagreement. We take the average score of these six sub-questions for every respondent, and calculate the top quartile score to be 5.667. A total of 154 respondents whose performance exceeds the benchmark are extracted from the dataset for further demographic analysis.

Figure 1 shows the age distribution of the loyal fans. People aged 45 to 64 accounted for around 50% of the loyal fans, followed by young adults aged 25 to 34 and elders aged over 65. *Figure 2* indicates the composition of loyal fans is nearly 60% male and 40% female.

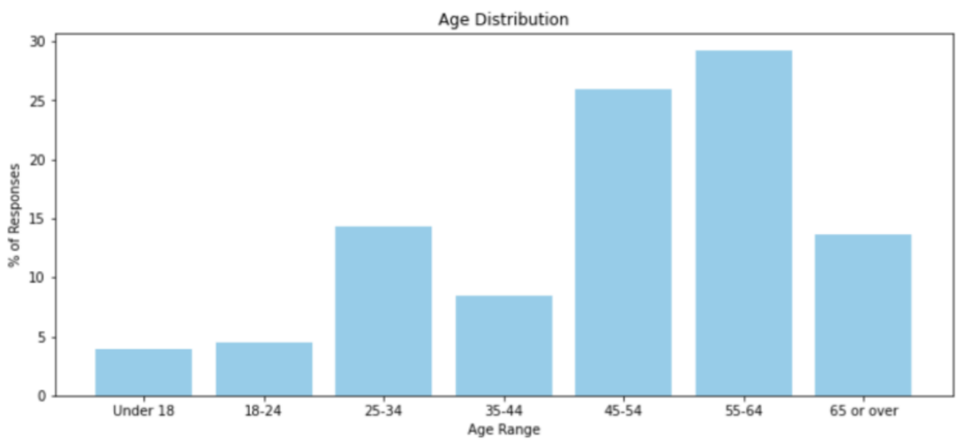

Figure 1. Age Distribution of loyal fans.

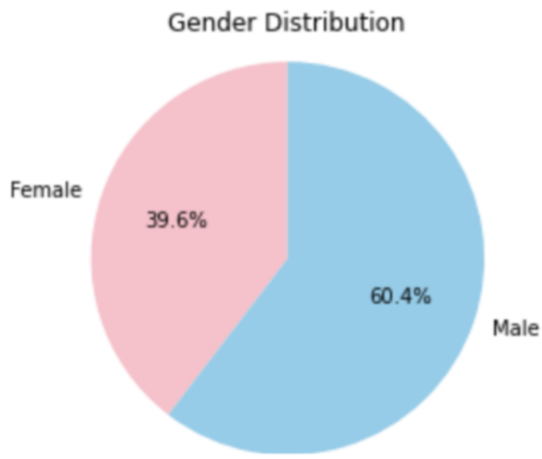

Figure 2. Gender Distribution of loyal fans.

Next, a binary random forest model is conducted to predict how likely these loyal fans would continue to support the Swans. A dummy variable is created on Question 14.1 where score 7 would be classified as 1 - definitely loyal and 0 - otherwise.

Figure 3 suggests that q142 and q145 are identified as important features in our model. Q142 has the highest score of 0.55, indicating that fans are keen to be a spectator at a game of Swans rather than any other sporting events. Q145 has the second highest score of 0.2, suggesting that no matter what, fans are devoted to Swans.

	Feature Importance
q142	0.549974
q143	0.048639
q144	0.058442
q145	0.204395
q146	0.138550

Figure 3. Feature importance of binary random forest model on question 14.

Several metrics are used to evaluate the performance of our model. As shown in *figure 4*, our model achieved an error rate of 0.178, indicating that it correctly classified 82.2% of the samples.

	Error rate	Precision	Sensitivity	Specificity	AUC	Cross-Entropy
Random Forest	0.177515	0.78022	0.876543	0.772727	0.906145	0.406171

Figure 4. Binary random forest model performance measures on Question 14.

Based on the results, as fans regard being a spectator of the game as very important, Swans could adjust their pricing strategy to incentivize loyalty among their members, for example developing loyalty programs.

Analysis 2

Questions	Target Variable(s)	Response Variable(s)	Model Name(s)
Who are the FWFs? What factors make them loyal to the Sydney Swan?	Q14	Q6 _ Q9	Logit Model

Fair-Weather Fans are the fans who have lower confidence and belief in the team and may change to support another team if the team does not have a good performance. To classify FWFs from the dataset, we add the sum of responses in question 14, and for a total score of 42, those whose score is lower than 30 is identified as the FWFs.

Next, a random forest model is implemented to find out the factors to predict FWFs' future support from question 6 to question 9. Question 6 and 7 are on-field issues, while question 8 and 9 are off-field issues. The model showed an accuracy of 0.72, with a confusion matrix of 76 true positives, 18 false positives, 20 false negatives, and 21 true negatives. The top ten most important features in the model are listed in *figure 5*, with q83 ("increase accessibility to the club"), q61 ("team gets results"), q64 ("play well under pressure"), q921 ("satisfaction with the game experience"), q85 ("provide good facilities to watch the game"), q81 ("the club management tries harder for members"), q62 ("play as a team"), q84 ("efficiently manage the club"), q912 ("offer exclusive benefits to members"), and q942 ("have a positive image of the club"). It is obvious that FWFs regard off-field issues more important than on-field issues.

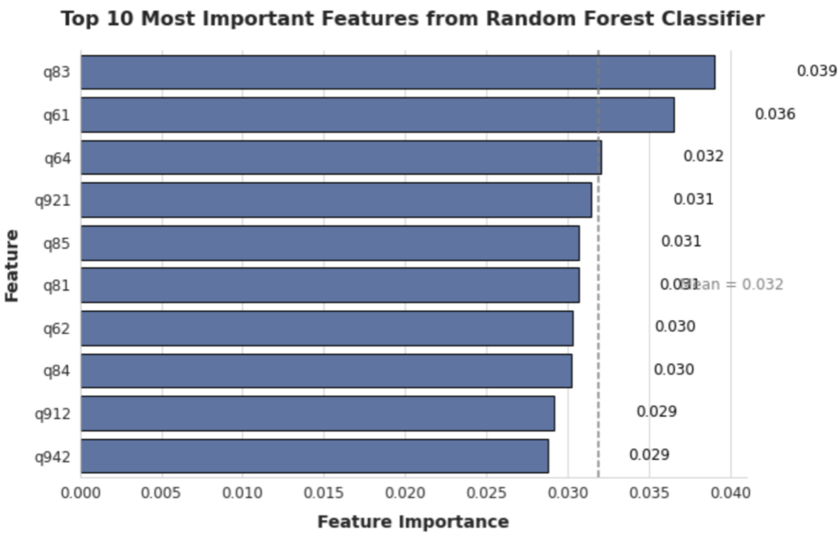

Figure 5. Random forest model performance measures on question 6 to 9.

Finally, a logit model is used to predict FWFs' status based on the ten most important features we identified above. The model yields an error rate of 0.38, with a sensitivity of 0.33 and a specificity of 0.8, indicating that the model correctly predicted FWF status 33% of the time and non-FWF status 80% of the time. The precision of the model was 0.52, which means that 52% of the predicted FWFs were FWFs. *Figure 6* summarizes the AUC score of all selected variables, suggesting that the model has moderate predictive power with an average score of 0.55.

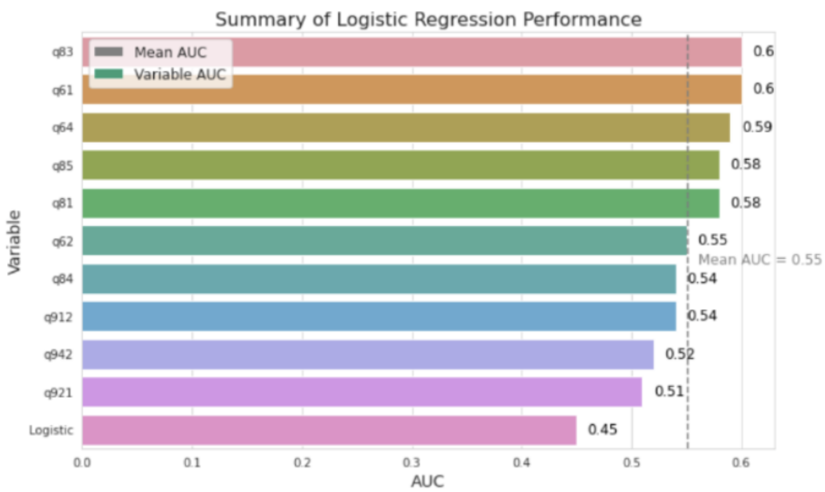

Figure 6. Logit model performance measures of the top 10 important features.

Based on the results, FWFs value the club's efforts to provide them with exclusive benefits and to increase their accessibility to the club as important factors to keep supporting the team(Coalter, F. 2007). Therefore, it is recommended that the management team focuses on improving the off-field performance of the team and fostering a strong sense of community among its fans. The club should consider providing better facilities and a visually appealing stadium to enhance the fan experience (Biscaia et al., 2016).

Analysis 3

Questions	Target Variable(s)	Response Variable(s)	Model Name(s)
What will affect overall members' future loyalty?	Q15	Q15.1 - Q15.10	Random Forest, SVM

Question 15 explores the factors that may influence an individual's decision to continue or terminate their membership with the Swans, such as changes in membership fees, game facilities, team performance, seat allocation, and even the team's location. The responses to these questions can help the Swans understand which factors are most important to their members and use that information to make decisions that prioritize member satisfaction and retention.

Figure 7 shows a strong relationship between q154 and q159 (0.89) and q152 and q155 (0.66), so they may require deeper analysis. *Figure 8* shows high dispersion for q154, indicating that members have varying opinions about game location.

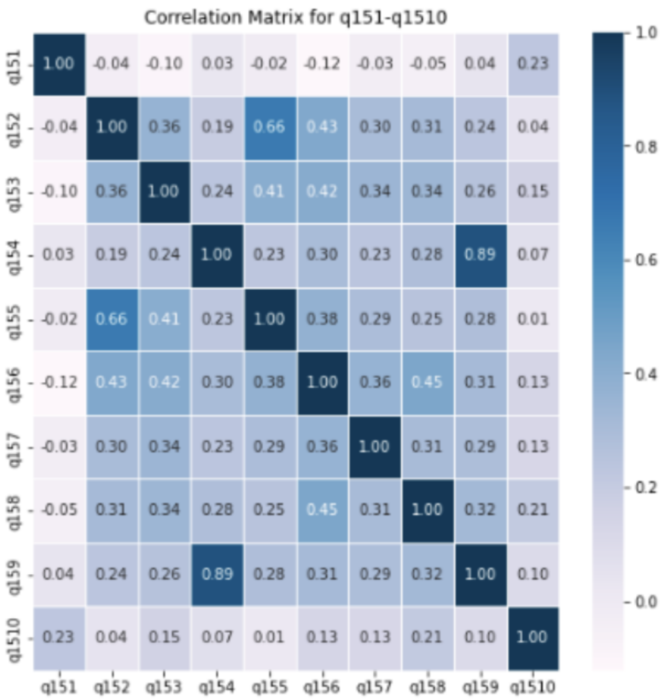

Figure 7. Correlation matrix of question 15.

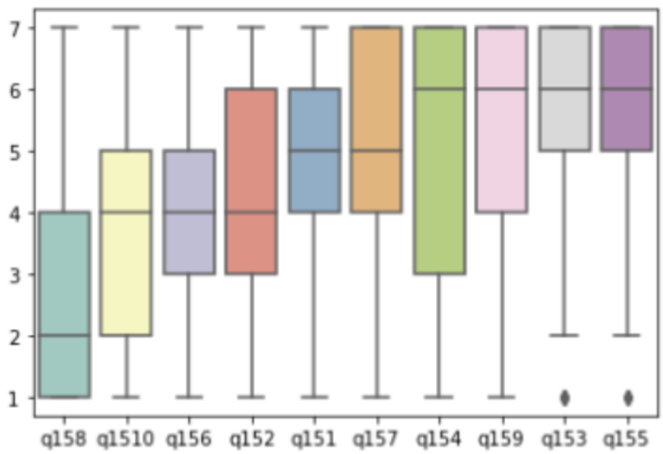

Figure 8. Boxplot of question 15.

To measure Question 15's overall performance, an average score of each participant is computed and a new column is created in the dataframe. If the mean of question 15 is above 4, it indicates member satisfaction with the opportunity to improve, otherwise, dissatisfaction. We opt for the SVM Continuous DV method as all the relevant data are continuous. Additionally, the features are interrelated and sequential, for instance, q152 and q155 relate to membership fees, while q154 and q159 relate to competition location. However, this model's drawback is its low transparency, and the intricate logic in the calculation process is concealed, leading to the "black box problem".

The SVR coefficient ratings below *(figure 9)* indicate that questions related to venues and facilities, particularly q154 and q151, hold the highest importance at 10% and 8% respectively. This is likely due to the significant role that venues and facilities play in the game experience for football fans. Following this are membership fees (q152 and q155) at 10% and 8% importance, which could impact Swan's pricing strategy, such as price skimming or penetration pricing. However, team performance (q153) and member benefits (q156, q157, q158, q159, and q1510) held lower coefficients, indicating that improving the game experience and keeping fees reasonable are the most critical factors in retaining Swans members, with team performance and member benefits following closely behind.

	SVR coefficient rating
q151	0.088078
q152	0.099056
q153	0.094422
q154	0.102689
q155	0.086263
q156	0.095169
q157	0.093203
q158	0.101887
q159	0.094086
q1510	0.095715

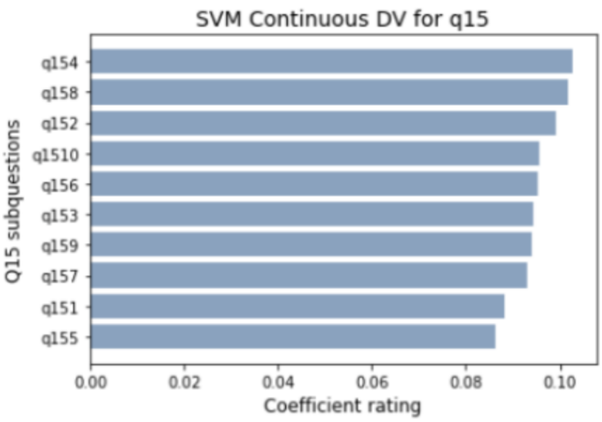

Figure 9. SVM model performance measures on question 15.

In order to measure the accuracy of the model, we establish benchmarks based on the Continuous Random Forest model to compare the following performances. As shown in *figure 10*, the MSE of SVR and Random Forest models for q15 is 0.002274, which means the average squared difference between the actual and predicted values is low. The MAE (mean absolute error) of both models for q15 is 0.038197, indicating that the average absolute difference between the actual and predicted values is low. The R-squared value for both models for q15 is high at 0.997773, indicating that the models fit the data well.

	MSE	MAE	RMSE	R-squared
SVR for q15	0.002274	0.038197	0.047681	0.997773

	MSE	MAE	RMSE	R-squared
Random Forest for q15	0.002274	0.038197	0.047681	0.997773

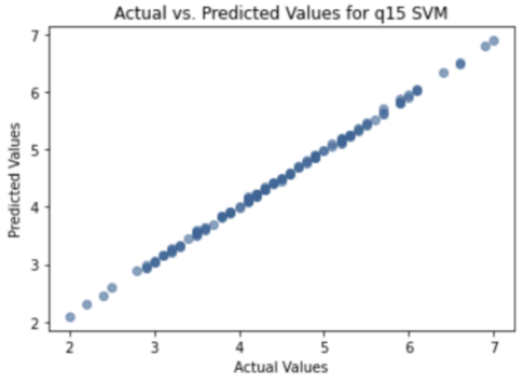

Figure 10. Model performance matrix on question15.

In sum, both SVR and Random Forest models for q15 show good performances with low error values and high R-squared values. However, the SVR model had higher accuracy and a stronger linear relationship between the predictor and target variables. This may benefit from the soft margin classification from SVM, which means the street can separate the positive cases from negative cases and limit margin violations while maintaining accuracy.

Based on the result, the management team could improve by considering changing the club location to Homebush as this is the most important factor that would influence current members' future support.

Analysis 4

Questions	Target Variable(s)	Response Variable(s)	Model Name(s)
Are those who attend games with families willing to pay 10% more in membership fees?	Q2.1 -Q2.10	15.5	XGBoost Model

As one of Swan's revenue streams is membership fee (Sean, 2016), we are interested in whether being accompanied by friends or families in a game would increase the willingness of paying more membership fees, 10% in particular.

Question 2 represents different status when attending a game, and question 15.5 refers to the likelihood that existing members are willing to pay 10% more for future membership.

An XGBoost model is implemented for prediction. *Figure 11* shows the feature importance on question 2, question 2.4 has the highest value of over 0.2, meaning members enjoy more when their family members and friends also attend the game. However, it is not significant enough to predict whether the members will pay an extra 10% or not because it only has 0.2 feature significance.

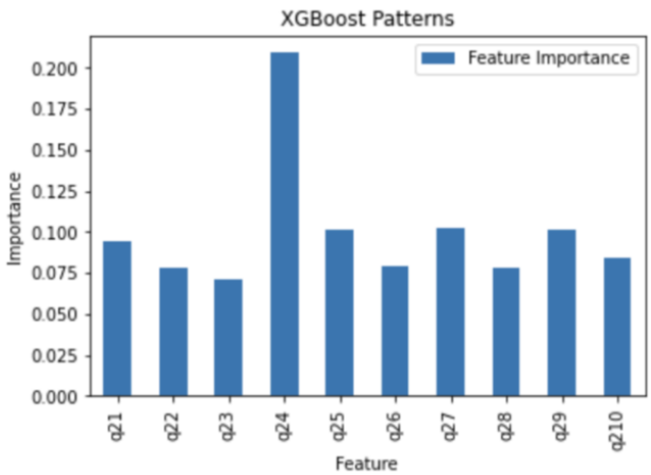

Figure 11. XGBoost model performance measures on question 2.

The accuracy rate of XG Boost is 0.84, which means 84% of the time the model is correct. The error rate is 0.19, which means it makes incorrect predictions for around 19%.

The force plot below *(figure 12)* shows how each feature affects the output of a machine learning model for a particular case. The length of question 2.4 has the largest value of 7 for the feature contribution to the output, which indicates that the corresponding feature has a relatively strong positive impact on the model's output. The color of the bar shows whether the value of the feature for the instance is more or less than expected. Blue bars indicate lower values, and red bars indicate higher values.

Figure 12. XGBoost model performance measures visualization on question 2.

Overall Recommendation

The management team of Sydney Swan could improve their business performance for the following aspects:

1. Adjust pricing strategy to incentivize loyalty among members. For example, offering discounts, developing loyalty programs and family packages.
2. Improve off-field performance (accessibility to the club and quality control of facilities) to enhance gaming experience.
3. Change the club location to another city, Homebush is recommended.

Limitation

Some questions in the questionnaire are not very relevant in the same section, which may lead to inaccuracy of the prediction. Also, transforming continuous data into discrete data, using Log and off for analysis may optimize the membership program.

Further Research

Conducting A/B testing or conjoint analysis on relevant variables could help identify the optimal game location for each customer segment (Wedel & Kannan, 2016). Effective communication with members is essential in improving their satisfaction and loyalty, and hosting member forums could help the team gain insights into the needs and preferences of their members (Gebert et al., 2017).

Further research and data collection could help the team identify new opportunities for growth and improvement, and enable them to make more informative decisions about their operations and strategy.

Reference

Biscaia, R. *et al.* (2016) "Investigating the role of fan club membership on perceptions of team brand equity in football," *Sport Management Review*, 19(2), pp. 157–170. Available at: https://doi.org/10.1016/j.smr.2015.02.001.

Coalter, F. (2007) "Sports Clubs, Social Capital and Social Regeneration: 'ill-defined interventions with hard to follow outcomes'?," *Sport in Society*, 10(4), pp. 537–559. Available at: https://doi.org/10.1080/17430430701388723.

Gebert, H., *et al*. (2017). Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts. Journal of knowledge management, 7(5), 107-123. *https://www.proquest.com/docview/230318587?accountid=14757&parentSessionId=eH0YCJZRIsjJmnFMRmfz411S8VZyShGx%2F8XPmxZN3l4%3D&pq-origsite=primo*

McKinsey & Company. (2017). Preparing for loyalty's next frontier: Ecosystems. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Marketing%20and%20Sales/Our%20Insights/Preparing%20for%20loyaltys%20next%20frontier%20Ecosystems/preparing-for-loyaltys-next-frontier-ecosystems.pdf

Wedel, M. and Kannan, P.K. (2016) "Marketing Analytics for Data-Rich Environments," Journal of Marketing, 80(6), pp. 97–121. Available at: https://doi.org/10.1509/jm.15.0413.