Sprint Challenge 3

Josh Gaughan

November 22nd, 2025

Question 1

For this question, I chose a peer-reviewed article which attempts to predict golf scores at the shot level. I will discuss how the article addresses each phase of the CRISP-DM.

To start, we will look at the business understanding phase. The writers of this article are not intending to serve a specific business requirement, but state in their abstract that: “This work helps players understand which skill sets they should improve on, manage courses better…and select the best tournament to enter.” (Drappi & Ting Keh, 2019) Their research is not driven by any stated needs of a company but seeks to create a new predictive model.

Next is the data understanding phase. In section 1.1, the authors identify the nature of relevant model work in the past. The previous regression models utilized aggregate factors to make a forecast of future scoring and earnings. The authors also discuss that the use of these correlated aggregate measures can produce undesirable results. Sections 1.2-1.5 recognize other existing metrics and useful research that will help in their analysis. In section 2, the authors discuss their contribution via the production of this model. In section 3, they highlight the ShotLink dataset as being the source of their data and the specific years being considered (2009-2016).

The data preparation phase involves constructing the data set that will be used in the models. In section 4 of the article, the authors list out the golfer, hole, course and game-state features that they used in their models. In 4.1.3, they: “...iteratively update time weighted features and normalized data until the difference of the two mean terms converges.” (Drappi & Ting Keh, 2019) They also talk about including weighting of past data as part of their analysis. In subsequent sections, they talk about the methodology behind the rest of the features that they use as part of their model. In some cases, they use raw features. In other cases, they use features they have modified. Some graphs and other visuals are provided throughout section 4 to better explain the nature of these features.

Section 5 discusses the model generation phase. In this section, they lay out the process of creating a benchmark by training neural nets and creating the forecasting model

using different methods. The training set consisted of the 2009-2015 season while the test set was 2016. Initially, they: “trained a hidden layer neural net of 50 nodes with golfer skill features described in section 4.1 and hole par” (Drappi & Ting Keh, 2019) before training a neural net with more features. They also tested out Softmax regression, Random Forest and a full feature Neural Network.

Section 5 and 6 contain results from the model. Of the models tested, the full feature Neural Network produced the lowest out-of-sample cross entropy error and appears to be the model they used. One example is: “In the 2016 Players Championship, on Jason Day’s R3, 15th hole 4th shot, the ball starts a few feet just outside the green. At that point the model predicted Jason’s score probabilities for the hole to be<par: 0.15, bogey: 0.81, double bogey: 0.03, double bogey +: 0.01>.” (Drappi & Ting Keh, 2019) The model provides a means to forecast score probability distributions leveraging the ShotLink dataset.

In section 6, they briefly discuss the potential use of the data fed into the model and the results generated from the model itself. They conclude with: “...these state variables are accessible in real-time, which motivates exciting applications. Applications that range from player development, course management, and tournament selection to audience engagement and improved sports books can easily be derived from this model.” (Drappi & Ting Keh, 2019)

The project overall appears to have useful results as it can provide knowledge to players and fans about the possible score outcomes for a player on a specific hole. If I were to provide any improvement to the article, I would suggest adding more visualizations. This way, readers would be able to better understand the model prediction tendencies. How much does the model favor certain player or game-state values? If a different player were asked to play the same shot, how differently would the model predict scoring outcomes? These types of questions could be answered by some additional visualizations or examples of model output.

Question 2

(2a)

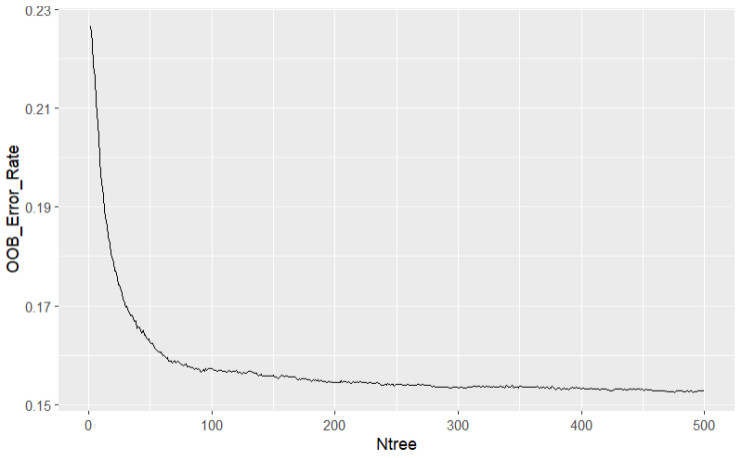
I ended up using data from 2021-2023 for the week 15 game between the Vikings and the Colts in 2022. The logistic model we discussed in class produced a classification rate of about 76% for my chosen data. In an attempt to improve on this, I tried several different models including a single decision tree classification model, lasso regression, feature engineering the predictors in the logistic regression we worked on in class and a random forest classification

model. In the end, the random forest model with tuned parameters produced the best classification rate at around 84.48% or about an increase of about 8.48%. (In my testing, this increase was fairly consistent across multiple different seasons. In some seasons/scenarios, the classification rate was closer to 90% for the RF model and the logistic was closer to 80%).

(2b)

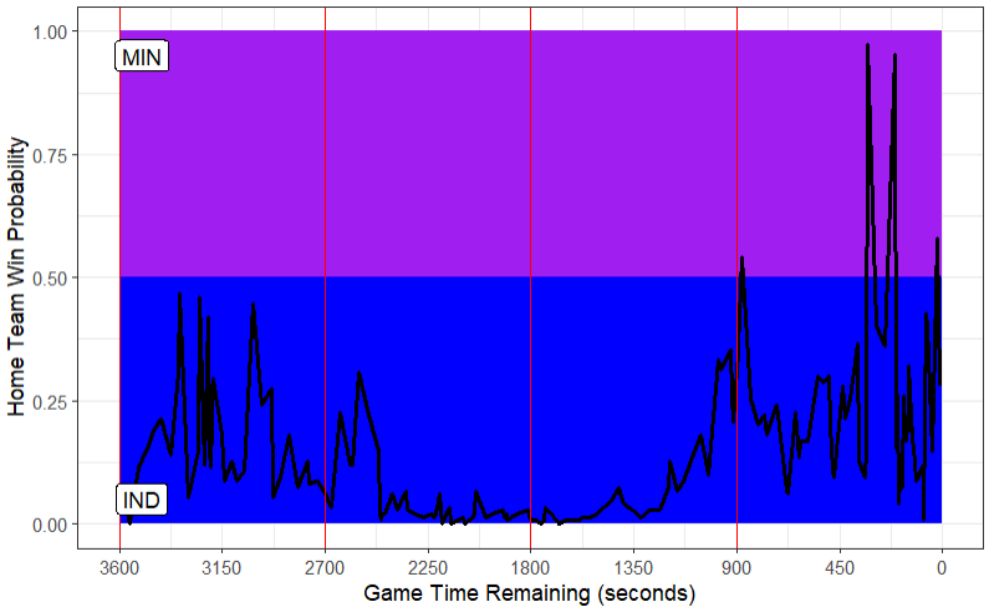
I wanted to try other forms of classification for this task. I tried some other methods and they didn't improve on the classification rate much. However, when I tried the random forest classification model, it showed a noticeable improvement in classification accuracy. I chose this model because it's more resistant to overfitting than a single decision tree and is an ensemble model which means it takes an average of several models which can give better results.

In order to make sure I got a well performing random forest model, I needed to make sure that I chose the correct number of trees (ntree) and number of predictors considered at each split (mtry) for the model. To find the optimal ntree value, I calculated the out-of-bounds error for several random forest models with different ntree values. As the graph suggests, we reach diminishing returns very early on around ntree = 150-200. I used 5 fold cross-validation to try and find the ideal number of predictors to consider at each split. Overall, I found that the higher the 'mtry', the lower the error rate. For this reason, I chose mtry = 6. This makes sense because there are strong predictors in our data and if mtry is low, there is a chance that these strong predictors are not chosen randomly during our splits. In the end, I used the 6 predictors from the original logistic model, ntree = 150 and mtry = 6 when constructing the random forest model. The overall classification rate is 84.48%.



(2c)

To the right, I have provided the win probability model for the 2022 week 15 game between the Minnesota Vikings and the Indianapolis Colts in Minneapolis. As we can see, the Colts have a high chance of winning to start despite being unfavored (spread) and away. This is most likely due to the poor calibration of the random forest model at the extremes even though the model predicts better than the logistic model. As you can see, the random forest maintains varying levels of confidence over 50% that the Colts will win the game, reaching nearly 100% at times in the middle of the game. In contrast with the logistic model, the random forest model starts to pick up on signs that the vikings may be making a comeback around the end of the 3rd quarter. The Colts' winning percentage according to the RF model at the end of the third quarter is around 70-75% whereas it is over 95% according to the logistic model. The RF model is much more sensitive to changes and is more likely to have jagged, extreme shifts in win probability. This is seen late in the fourth quarter where the Vikings tie the score and appear to be making the upset but the model is still uncertain who will actually win.



What's interesting about this situation is that, while the model has better predictive accuracy, it is more confusing to look at on the graph than the logistic model. What is also interesting is that, games like this where the Vikings come back from a 33-point deficit, are the edge cases that most predictive models are going to struggle with. Most, if not all models will have the Colts at or near 100% in the middle of this game because making up a 33-point deficit is seen as impossible. I am throwing predictive models at this incredibly unlikely game and

hoping that it predicts correctly. It isn't a surprise that it struggles with predicting the winner in the end.

Question 3

(3a)

There are a couple of important descriptive results that are important to the business perspective.

First, with the logistic model I created (see 3b for more information), we can look at the magnitude and sign of the coefficients to understand which features the business may choose to

	Estimate <dbl>	Std_Error <dbl>	Z_Value <dbl>	P_Value <chr>
Most_Frequent_Stadium_LevelUpper	-1.20969124	0.026762450	-45.201065	< 1e-10
Playoff_Appearance	-0.93085731	0.024461131	-38.054550	< 1e-10
total_num_non_resale	0.06788174	0.003736660	18.166422	< 1e-10
Most_Frequent_ClassGA	0.48392976	0.044044960	10.987177	< 1e-10
total_num_non_resale_scanned	-0.04057929	0.004044982	-10.032008	< 1e-10
Most_Frequent_Field_ViewMidfield	0.46089787	0.046463149	9.919643	< 1e-10
avg_resale_markup	-0.13732327	0.014484173	-9.480919	< 1e-10
total_num_seats	-0.02125351	0.003055587	-6.955621	< 1e-10
Most_Frequent_Field_View20-40	0.15183767	0.040187959	3.778188	0.00015797
Most_Frequent_Stadium_LevelMiddle	-0.11781854	0.038857656	-3.032055	0.00242895

focus on. Most_Frequent_Stadium_LevelUpper is listed as the most significant variable. This means that upper level ticket buyers have much lower log odds of retention. This may indicate that the tickets are too expensive for the fans that buy from this section so they are less likely to return. Adding some sort of incentive or lowering the price may increase retention. When all the variables were present in the model in my testing, Win_Pct, Hwin_Pct and Playoff_Appearance had very odd coefficient values (large in magnitude). After some research, I realized that the issue may be due to multicollinearity between these variables. I tried removing Win_Pct and Hwin_Pct and it improved the magnitude of the Playoff_Appearance coefficient but flipped the sign in a direction that was a bit confusing to me. The model indicated that a playoff appearance actually decreased the log odds of retention rather than increased it. This variable likely needs to be investigated further. The total_num_non_resale variable is also quite significant. This may indicate that customers who purchase big ticket packages and don't resell those tickets are loyal customers. The table also displays several other significant variables which provide additional descriptive results.

For the logistic model, I have also provided the confusion matrix. This displays the classification performance of the model. While we are able to identify the classification rate as being 79.695%, we are also

	Reference	
Prediction	0	1
0	750	1090
1	1360	8866

interested in the misclassifications. The false positive results (1360), are customers that were predicted to renew or be retained but who actually did not. These customers were overlooked by the typical efforts of customer retention and are the most valuable targets. The false negatives (1090) are the customers that were predicted to not be retained but actually were. These customers were focused on when it came to retention efforts but the resources used on them were wasted.

(3b)

For this question, I decided to stick with the same predictors we used in class but switched to a logistic regression model instead. This means that the model prioritizes interpretation over predictive accuracy. I ran into a few issues while trying to create this model and made some changes to try and mitigate them.

First, the confusion matrix from the random forest model we created in class displayed a large amount of false positives and a class imbalance. To fix this, I tried implementing class weights to the model. The intention was to assign greater weight to the minority class and lesser weight to the majority class during model training.

Second, I needed to adjust the classification threshold. This way, I'd be able to find a good classification rate while also reducing the false positives rate (increasing the true positive rate / precision). The random forest model had a precision of 83.4% while my logistic model had a precision of 86.7%. However, it is important to note that while the false positive (predicted to be retained but actually were not) rate decreased the false negative (predicted to not be retained but actually were) rate increased.

Finally, I needed to remove Win_Pct and Hwin_Pct from the logistic model in order to provide more realistic coefficients. There appeared to be some multicollinearity between these two variables and Playoff_Appearance which was drastically increasing the magnitude of the variables.

In the end, I fit a logistic regression model with all of the same variables as the random forest except for Win_Pct and Hwin_Pct. I applied class weights and made the classification threshold 0.35 instead of 0.5. This produced a classification rate of 79.695% versus the classification rate of 82.58% of the random forest model. The false positive rate went from 83.4% in the RF model to 86.7% in the logistic model. Sensitivity went from 1.54% to 10.948%. However, the F1 score (which is a harmonic mean of precision and sensitivity and is defined as better when it is higher) of the logistic model is 87.86% versus 90.3% for the random forest. So, even though we have a higher classification rate and a better F1 score for the random forest

model, it isn't a massive improvement. Since the logistic model is more interpretable, it may be more helpful depending on the specific context.

(3c)

In the short term, business personnel should focus on minimizing wasted resources (on false positives). To do this, they should allocate a greater share of resources to this group since money spent elsewhere may be less necessary to retain a customer. However, it is important to not allocate all of the resources to this group, just more. Business personnel should also consider adding incentives for upper level customers since they are a group that is less likely to be retained. Additionally, personnel should also address customers with high resale activity and markups by potentially adding incentives that limit their resale activity.

In the long term, the business may consider finding ways to provide greater rewards for buyers who commit to more expensive non-resale packages. There should also be a review of the ticket pricing strategy for upper level seats. Customers may be finding the seats to be overpriced for what they offer which could be causing the decrease in retention. Focusing on improving the value for these seats may change the most significant variable coefficient Most_Frequent_Stadium_LevelUpper from very negative to less negative or even positive.

Works Cited

Drappi C., Co Ting Kek, L., (2018). Predicting golf scores at the shot level. Journal of Sports Analytics, https://journals.sagepub.com/doi/full/10.3233/JSA-170273