# INFSCI 2160 Project Progress Report

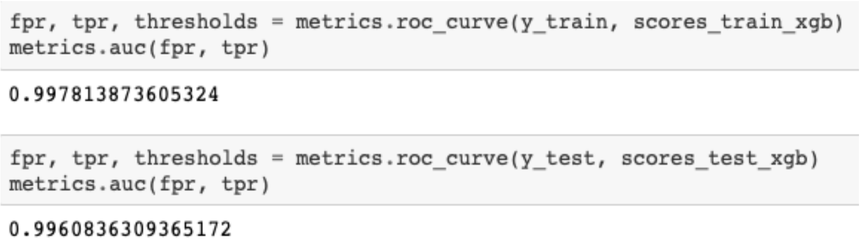
**Group member:**

Tshering Sherpa, Yanru Li, Yuyan Li

For our project we were working on the Prediction of Heart Disease from the data called Heart Disease Data Set from the UCI Machine Learning website (https://archive.ics.uci.edu/ml/datasets/heart+Disease). There were only 300 rows of data so we decided to change our data source. Now we are working on the new dataset. In this project, we use the dataset “FIFA 19 complete player” from kaggle. FIFA stands for the Fédération Internationale de Football Association which is a non profit organization which describes itself as an international governing body of association football. In the USA it is called soccer. Our dataset is the latest edition of FIFA, and the data is scraped from https://sofifa.com/.

From this data set we chose the features that we thought was important for our project. We made the value of the player as a target and changed it to binary so that we can find out if the value of player will be higher than 100 million euros or less. We split the data in train and test set and use XGboost, lightgbm, Catboost, and logistic regression models to get our auc score. To get the better results we did grid search and hyperopt. We then did shap and interaction plot to find out the most important features from our data.

After running through our code multiple times we got the same results with 0.99 auc score for XGBoost, lightgbm, and Catboost models. Figure 1 shows the result for XGBoost, which auc for training data is 0.998, and auc for testing data is 0.996.



```
fpr, tpr, thresholds = metrics.roc_curve(y_train, scores_train_xgb)
metrics.auc(fpr, tpr)
```

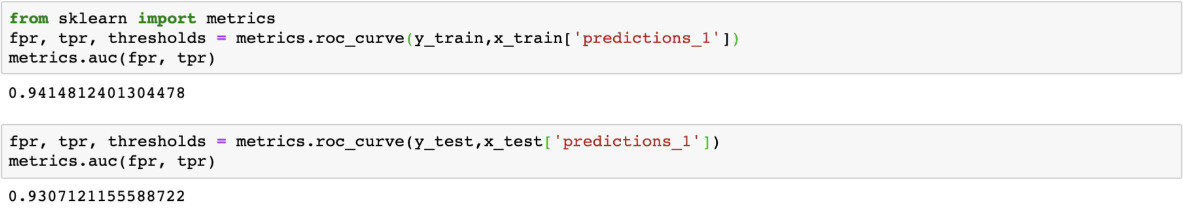
```
0.997813873605324
```

```
fpr, tpr, thresholds = metrics.roc_curve(y_test, scores_test_xgb)
metrics.auc(fpr, tpr)
```

```
0.9960836309365172
```

Figure. 1

For the logistic regression model, we got the auc score for training set is 0.94 and for test set is 0.93. Figure 2 shows the results for logistic regression model.



```
from sklearn import metrics
fpr, tpr, thresholds = metrics.roc_curve(y_train,x_train['predictions_1'])
metrics.auc(fpr, tpr)
```

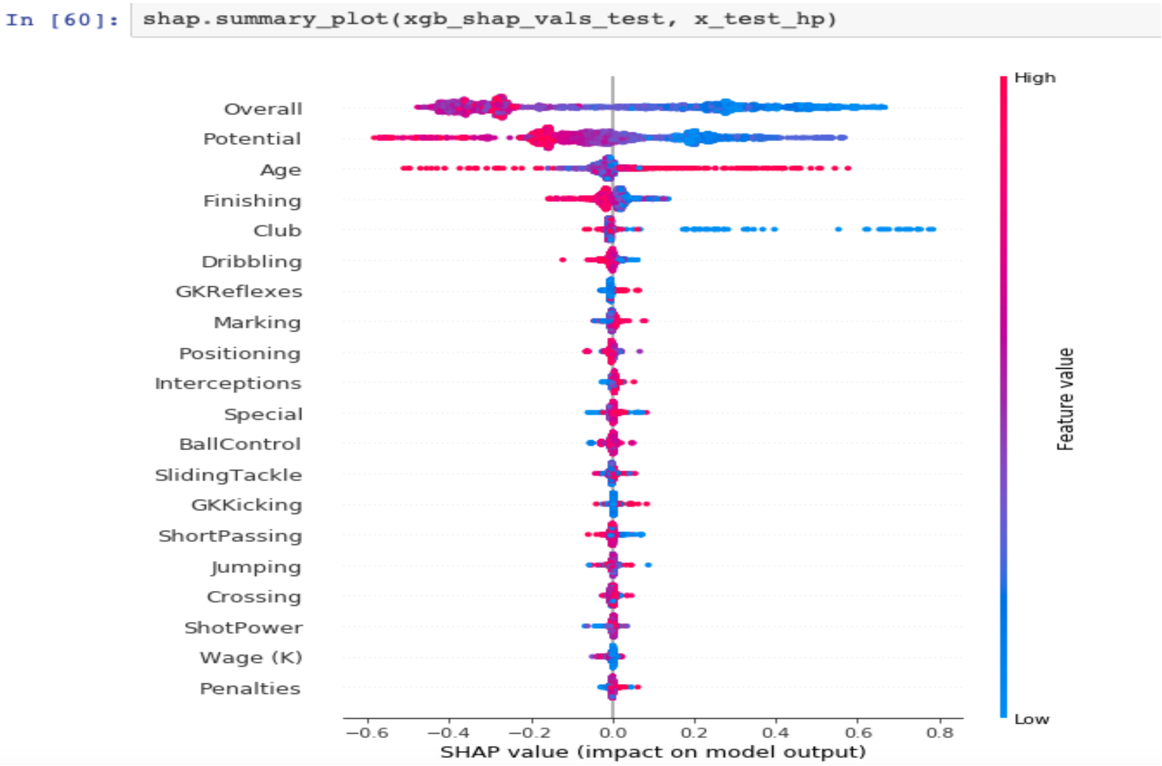
```
0.9414812401304478
```

```
fpr, tpr, thresholds = metrics.roc_curve(y_test,x_test['predictions_1'])
metrics.auc(fpr, tpr)
```

```
0.9307121155588722
```

Figure. 2

This tells us that our models fit very well on our data. Figure 3 shows the shap summary plot, which gave us that age is the most important feature from our data.

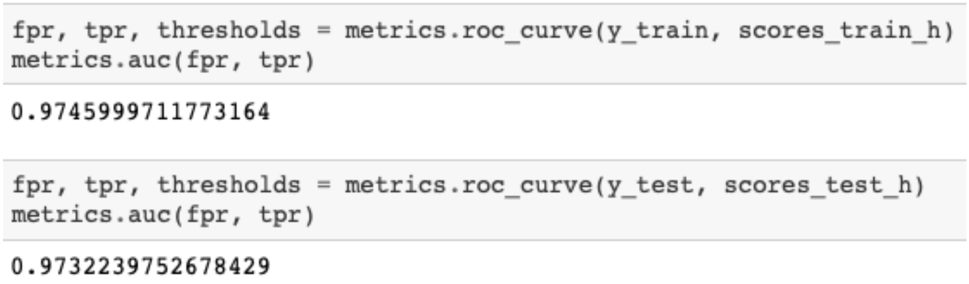


```
In [60]: shap.summary_plot(xgb_shap_vals_test, x_test_hp)
```

Figure. 3

FIFA is multi-billion company and if we can predict the value of player then we can let FIFA know what the player is worth and how other clubs can deal on buying or trading him/her with different clubs. The players are often traded or bought from one club to another and the clubs pays almost more than 100 million euros. Therefore it is crucial that the clubs know what’s the actual player worth.

Even though we use hyperopt to get better results, we ended up getting lower auc scores. This might be the results of our models. Figure 4 shows the r esult for hyperopt, which auc for training set is 0.975, and auc for testing set is 0.973.



```
fpr, tpr, thresholds = metrics.roc_curve(y_train, scores_train_h)
metrics.auc(fpr, tpr)
```

```
0.9745999711773164
```

```
fpr, tpr, thresholds = metrics.roc_curve(y_test, scores_test_h)
metrics.auc(fpr, tpr)
```

```
0.9732239752678429
```

Figure. 4

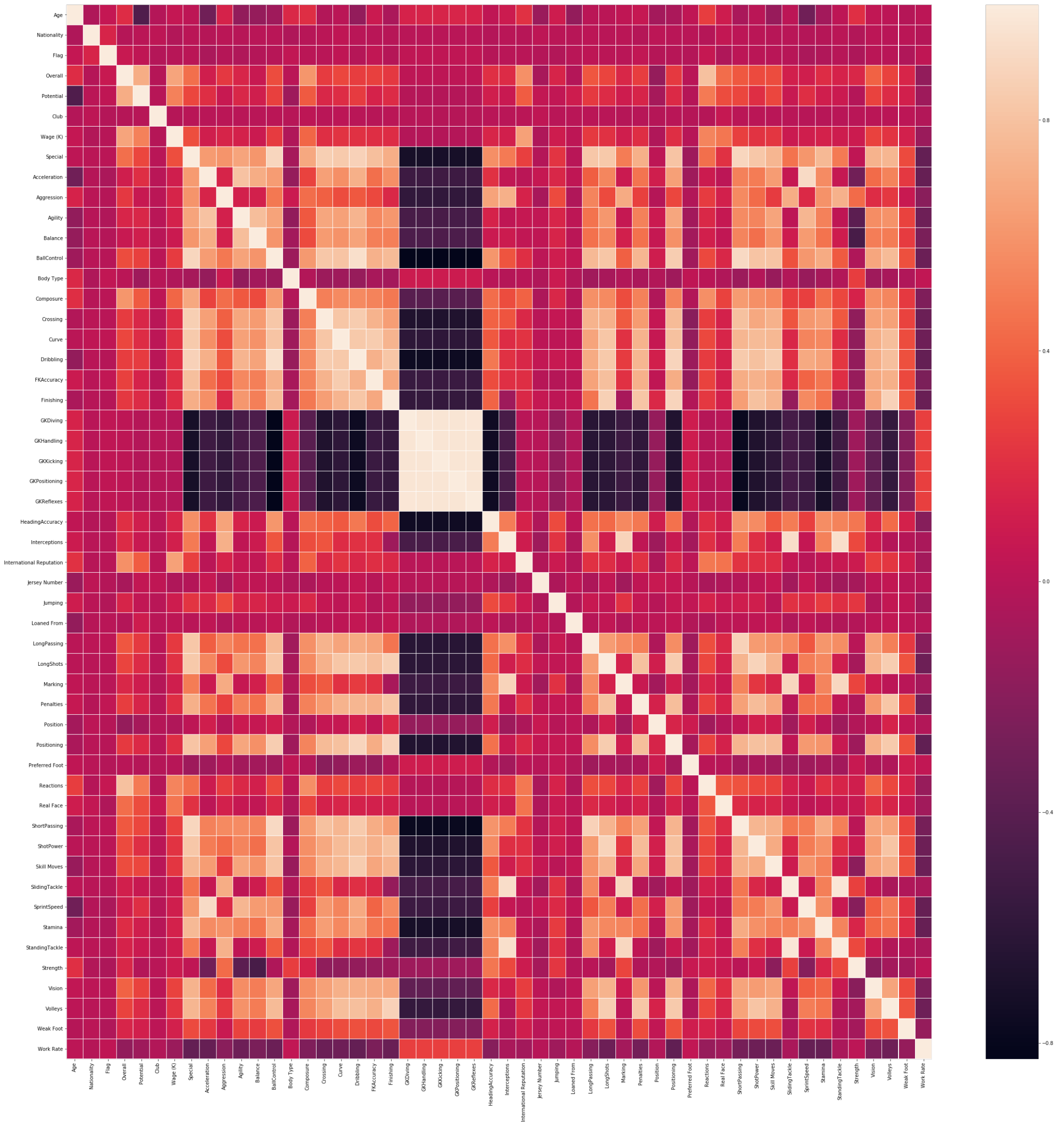


Figure. 5

We thought the reason all models appear very good with auc 0.99 is the variables are highly correlated with each other, which makes the model very good. Thus, we did the heatmap (Figure 5) for the variables we have used to cl assify. From the heatmap, we can easily see the correlation between variables, which black means variables have highly negative relationship, and white means variables have highly positive relationship. It makes sense since the dataset is about the scores the players get in each exercise, and players would get a higher score if they are great.

We have selected all exercises to be our variables, and most players have very high grades for these exercises, it makes our model too good to predict the value players worth. Since our results were unexpected, we will keep working on to get well fit models. Thus, in the next step, we want to reselect the variables, to filter similar variables to make the model cogent, and to run through our code again. Also we are planning to use lasso and ridge algorithm to see if we will get better results.