Identifying “Gold Players” using FIFA 2018 DATA

*Lijia **LIU***

*Handong **HOU***

*Jinyi **FANG***

*Jinyong **MA***

*Aviroop **GHOSAL***

*Abhinaw **PRIYADERSHI***

*Jie **ZHAO***

A. Executive Summary

Project Description:

In this project, our group aspires to create an effective model that identifies the most profitable FIFA players, so that clubs can acquire high-performing players at a fair cost and enjoy a great value from them in return. In this way, we can help clubs dramatically increase their profits and give them a competitive advantage in future.

Our approach is to first thoroughly understand the business problem, and then design a process flow to solve the problem. We began by acquiring in-depth external data from credible online sources such as Kaggle. Once we had all the relevant data, we then cleaned the data by taking out data points with any error and null values, as well as entries that do not carry important business meaning in regards to our objective. After cleaning we merged a data set to bring down the club numbers from 740 to top 100.This helped us in performing profit-loss analysis per player .Then, we created a Logistic Regression model and an Ensemble model to predict the most profitable players more accurately.

In the end, we achieved great success from both business and statistical perspective. We are able to help the clubs acquire high-value players at a low cost. Therefore, we can greatly reduce the clubs' loss or increase their profit. Additionally, we can help the clubs form strong teams that have a high chance of winning future games. In terms of the model success, the Key Results section includes all relevant model performance.

Key Results:

We selected a total of seven variables: Acceleration, Reactions, Positioning, GK reflexes, Free kick accuracy, Ball control, Composure, which were selected based on model significance and business knowledge. We first built a logistic regression model with R square of 57.66% for training and 63.29% for testing. At the same time, the accuracy for training is 90.45% and for testing is 91.38%. Then we created our best model, which was an ensemble model of Logistic Regression, kNN, Gradient Boosting and AdaBoost. We achieved 94.6% accuracy on the Training set and 92.5% on the Testing set. The Training R2 was 69% and the Testing R2 of 60%. By applying our best model, we can help the clubs increase their profits by €0.57 million per player.

B. Table of Contents

C. Introduction

Background and Motivation: It's the FIFA transfer season and club managers are concerned about getting the best players with high-performing players and add to their existing team that enhances average profit of the club per player. We, as the experts in a consulting firm, are highly motivated to render advisory service to help the clubs target good players ("Gold" players), and then increase the profitability of the club by acquiring a certain number of golden players.

Objective: Increase profits for the FIFA clubs by identifying the most profitable players for the club.

D. Problem Statement and Hypothesis

Problem: The FIFA transfer season is stressful for all the club managers since they are not only concerned about finding a high-performing player, but also ensuring that he is profitable to the team. A majority of club managers are not able to identify key profitable players for their teams. How can we, as a FIFA consulting company help the clubs enhance profits?

Primary Hypothesis: Players, typically within an age group of between 21 and 29, who have better ball control and acceleration are not only highly valuable but also contributing to significant profits to the club.

E. Procedure and Methodology

To prove our hypothesis, we performed the following steps:

Step 1 - Data Collection & Merging: We collected our main dataset**,** dataset of soccer players, from kaggle.com. (source: *Kaggle FIFA 18 Demo Player Dataset*) and merged it with external dataset, market value for top 100 valuable FIFA clubs. (source: *FIFA Top 100 Most Valuable Clubs*)

Step 2 - Data Cleaning: As per merging we removed clubs which valued greater than 100. Data is reduced significantly. After that we further cleaned our data by removing blanks, null values, missing cells and inconsistent data points.

Step 3 - Data Exploration: We transformed our "Overall" variable to a binary variable that we used as our response variable. This was further named as "GOLD_Overall". We then identified 3 variables (age, acceleration, ball control) from our business intuition that would affect the manager's decision in selection of the most profitable players, but building a logistic regression model with these variable, gave us R2 of only 16%, and thus we further exploring other significant variables. We looked at all the variables and based on our domain knowledge, we categorized them into **Demographics** (Age, Nationality, Club, Wage) and **Techniques** (acceleration, ball control, reactions, crossing, curve, dribbling, etc..) variables.

Step 4 - Business intuition on important variables: From our business intuition, we identified 3 variables to be strong predictors for Gold players. These variables are Age, Acceleration and Ball control. Firstly, it is easy to understand that Age is supposed to be significant since either the older players have accumulated professional experience or the younger players keep a more athletic condition. Secondly, as for techniques, we believed that if the players are good at acceleration and ball control since faster players or players with good ball control, they will be more likely to become Gold Players.

Step 5 - Identify Significant Predictors using Stepwise Regression and DOE: Running Stepwise Regression, we targeted 11 significant predictors for following analysis, including Wage, Acceleration, Balance, Ball Control, Composure, Finishing, GK reflexes, Heading accuracy, Long Passing, Positioning, and Reactions. Subsequently, we dug deeper for DOE analysis according to the significant variables we discovered. Although there are some new variables significantly related to Gold Payer, it does not make business sense. Fig [1], Fig [2].

Step 6 - Exploratory Data Analysis: We looked at the correlation and pair plots to check for strong and weak correlation among our predictors with our response variable. In addition to the important variables, from our intuition we concluded that ***Reactions*** and ***Composure*** were also good predictors to predict a Gold player. This was logically true because good players can be not only rated high on their stats, but also strongly reactive to the situations around them. Great players also tend to be very calm and composed. We decided to further dig into these player attributes and plot a histogram plot look at distinguish characteristics between “Gold” and “Not Gold” players. We noticed that the Gold players tend to rate higher on all these attributes - Acceleration, Free Kick Accuracy, Positioning, Composure, Reactions, Ball Control. This made sense because a very valuable player would be fast, score free kicks frequently, position efficiently to receive passes, have good ball control. Fig [3], Fig [4], Fig [5].

Step 7 - Build Model: Model Building and Comparison: Our best model was an ensemble model of Logistic Regression, Gradient Boosting and AdaBoost Network. The input variables used in the all the models were same - Acceleration, Reactions, Positioning, GK reflexes, Free kick accuracy, Ball control, Composure. We ran the below mentioned models on our training data and attained the following Training and Testing accuracy scores.

Step 8 - Profit/Loss Analysis: We defined the profit or loss by subtracting the player value (Since player value indicates how much the club paid for the player) from the average revenue of a player in a club, which comes from our merged data set : **Profit = Revenue (Average Revenue of a player per club) - Cost (Player value).** To calculate the profit and loss, we only included the profit and loss from the gold players. So we are comparing the difference between the average profit/loss from predicted Gold players for each club and average profit/loss from the original Gold players. The final output for the profit is the average of each club’s average profit/loss per player.

F. Data

ACCELERATION: Acceleration means the rate of speed at which players move, and this is one strong attribute to judge a player's game skill. There are outliers, which we assume are valid. It shows a left-skewed distribution. The mean and the median are 66.9 and 69 respectively. The standard deviation is 14.8 and the range is 83, varying from 13 to 96. Fig [13]

BALL CONTROL: Ball control means the capability to control the soccer ball, and this is one criterion for players. There are outliers, which we assume are valid. It shows a left-skewed distribution and. The mean and the median are 64.8 and 70 respectively. The standard deviation is 18.4 and the range is 85, varying from 10 to 95. Fig [13]

FREE KICK ACCURACY: Free kick means an unimpeded kick of the stationary ball awarded to one side as a penalty for a foul or infringement by the other side. There are not outliers and it shows a normal distribution. The mean and the median are 47.7 and 48 respectively. The standard deviation is 19.5 and the range is 88, varying from 8 to 96. Fig [13]

POSITIONING: Positioning means an overall score for being forward, midfielder, defender, or goalkeeper. There are not outliers and it shows a normal distribution. The mean and the median are 55 and 61 respectively. The standard deviation is 21.7 and the range is 92, varying from 3 to 95. Fig [13]

REACTIONS: Reactions mean the comprehensive ability of players reacting to the contest situation and emergencies in the field. There are a few of outliers, which we assume are valid. It shows a left-skewed distribution and. The mean and the median are 68.7 and 70 respectively. The standard deviation is 10 and the range is 66, varying from 30 to 96. Fig [13]

G. ETL (Extract, Transform, Load)

How we got the data: For the main dataset, we found an interesting dataset for us to analyze, that is, a dataset of soccer players from a computer game named FIFA 18 including 35 attributes through kaggle.com. (source: *Kaggle FIFA 18 Demo Player Dataset*)For the external dataset, we find the market value data for top 100 valuable FIFA clubs on an website. (source:*FIFA Top 100 Most Valuable Clubs*)

How we subset the data: As there are players for over 700 clubs in the original dataset from kaggle, we extract the players for the top 100 clubs correspond with the external dataset. We only want to build prediction and profit-loss model for the top 100 valuable clubs.

How we merge the data: We merged the FIFA players dataset with average market value of players for the top 100 valuable clubs. So we add the new column called average market value from the top 100 clubs data to the subset of FIFA players data.

How we cleaned the data: Since there were many inconsistent data points in players' performance attributed with extra bonus points added to some values like "56+3", "60-3", due to which the scores were string values and not numerical. We added these bonus points to base score by writing a Python script and further converted them into a numerical score.

How we made "Y "variable:We transformed the "overall" variables into our response variable: **"GOLD"** by categorizing the players with overall score larger than 79 (Since 79 is the upper quartile or the 75th percentile score - Thus indicating those players having exceptionally high overall scores lying the in the higher end of the spectrum) as a gold player, labeled 1 and the overall score less than 79 as the opposite, labeled 0.

Training and Testing: We split the data into training and testing with the ratio of 1:1.

H. Analysis

Second Best Model - Logistic Regression Model

Our second best model is a logistic regression model with 7 variables: Acceleration, Reactions, Ball control, positioning, GK reflexes, composure, Free kick accuracy. We selected these variables first by choosing the ones that make business sense and then using stepwise to select significant variables: Acceleration and Ball control. This logistic regression model with R square of 59.06% for training and 61.72% for testing. At the same time, the accuracy for training is 90.45% and for testing is 91.38%. Fig [7], Fig [11]

Best Model - Ensemble Model

Our best model is an Ensemble model with 7 variables: Acceleration, Reactions, Ball control, positioning, GK reflexes, composure, Free kick accuracy. We achieved a Training accuracy of 94.6% and Testing accuracy of 92.5%. The best model had a training R square of 69% and testing R square of 60%. This model improved the accuracy by around 1% on the Testing data. This model was more reliable and versatile since it avoided any overfitting issue and can work more accurately on a new dataset.

Data Analysis: We looked at the descriptive statistics of the 5 important variables using JMP: Acceleration, Ball Control, Free Kick Accuracy, Positioning and Reactions and noticed that most of them were left skewed normal distribution(Fig [13]). Using Python, we built a correlation plot (Fig [3]) and pair plot (Fig [4]) to understand the relationships between our predictors and check for any multicollinearity issues that would affect regression models and also find relationships between the predictors and probability of a Gold player. Through this, we further found out Reactions and Composure

are also strong predictors to predict probability of a Gold player.. Then, we plotted variable distributions to distinguish between our "Gold" and "Not Gold" players. (Fig [5]). We noticed these "Gold" players can be distinguished from others in the performance KPIs such as Positioning, Free Kick Accuracy, Reactions, Composure, Ball Control. This indicates these Gold players rate higher on their skill and technique metrics.

Model building: We built a Logistic Regression model on JMP using significant predictors as mentioned above and attained a high accuracy on both Training and Testing set with only a difference of 1% between them. Then we used Python Scikit-learn to build new models. We noticed that as we tried out new models, the model did a great job learning the data, so it was difficult to understand whether the improvement in score came from capturing the relationships more accurately or overfitting the data. Hence, we performed *Stratified k-fold Cross Validation* on our training data to divide the data into 10 folds. Then, we kept a holdout sample and trained the model on the remaining data. This helped in analyzing the effectiveness of a model's performance. (Fig [6])

We attained the following Cross Validation mean scores to observe the most effective models.

Models	Training Accuracy	Testing Accuracy
Logistic Regression	90%	91%
kNN	92%	90%
Decision Tree	95%	87%
Random Forest	96%	91%
Gradient Boosting	96%	92%
AdaBoost	93%	91%
Neural Network	77%	76%

Table[1]: Model Cross Validation Mean Accuracy comparison

We combined our interpretation from the graph and observation from our initial run of model training and testing accuracy, we selected Logistic Regression, kNN, Gradient Boosting and AdaBoost to make an Ensemble model. These models did not overfit the data and had more accurate predictions.

I. Performance Measure

Second Best Model-Logistic Regression

Training Confusion Matrix:

		Predicted		
		Not Gold	Gold	
Actual	Not Gold	TN=966	FP=46	1012
	Gold	FN=79	TP=219	298
		1045	265	1310

Testing Confusion Matrix:

		Predicted		
		Not Gold	Gold	
Actual	Not Gold	TN=968	FP=58	1026
	Gold	FN=55	TP=228	283
		1023	286	1309

Training Accuracy: 90.45% (The model has predicted 90.45% of the total training data accurately)
Testing Accuracy: 91.38% (The model has predicted 91.38% of the total training data accurately)
Training R2: 57.66% (For the training data, around 57.66% of the variability of the "Gold" variable could be explained by the input variables)
Testing R2: 63.29% (For the testing data, around 63.29% of the variability of the "Gold" variable could be explained by the input variables)
AUC: This is the area under the ROC curve. The higher the better, hence it has classifies positive samples and negative samples quite accurately.
Training AUC 0.97 - Fig [8]
Testing AUC: 0.96 - Fig [12]

ROC Curve - The ROC Curve is plotted with varying thresholds. It is towards the true positive rate rather than false positive rate. This implies that the majority of positive and negative samples in testing data has been classified accurately.
Training ROC Curve - Fig [8]
Testing ROC Curve - Fig [12]

Nominal Logistic Statistics, ROC Curve, **Lift Curve** for training are attached. Fig [7], Fig[8].

Best Model - Ensemble Model

Training Confusion Matrix:

		Predicted		
		Not Gold	Gold	
Actual	Not Gold	TN = 1000	FP = 12	1012
	Gold	FN = 58	TP = 240	298
		1058	252	**1310**

Testing Confusion Matrix:

		Predicted		
		Not Gold	Gold	
Actual	Not Gold	TN = 991	FP = 35	1026
	Gold	FN = 63	TP = 220	283
		1054	255	**1309**

Training Accuracy: 94.6% (The model has predicted 94.6% of the total training data accurately)
Testing Accuracy: 92.5% (The model has predicted 92.5% of the total testing data accurately)
Training R2: 69% (For the training data, around 69% of the variability of the "Gold" variable could be explained by the input variables)
Testing R2: 59.8% (For the testing data, around 60% of the variability of the "Gold" variable could be explained by the input variables)

AUC: This is the area under the ROC curve. The higher the better, hence it has classifies positive samples and negative samples quite accurately.
Training AUC 0.99 - Fig [9]
Testing AUC: 0.97 - Fig [10]

ROC Curve - The ROC Curve is plotted with varying thresholds. It is towards the true positive rate rather than false positive rate. This implies that the majority of positive and negative samples in testing data has been classified accurately.
Training ROC Curve - Fig [9]
Testing ROC Curve - Fig [10]

To summarize, we have the following table for training and testing for first and best models:

Model	Second Best		Best	
Name	Logistic Regression		Ensemble Model	
Parameters	Training	Testing	Training	Tesing
Accuracy	90.45%	91.38%	94.60%	92.50%
R2	57.66%	63.29%	69%	59.80%
AUC	0.97	0.96	0.99	0.97

Table[2]: Summary of Training and Testing for Models

J. Business Insights

From business perspective, a soccer club manager has two main roles for the club. First, the manager is supposed to bring more excellent players to the team to win maximum games in the season. Second, the manager must have an insight to bring players with high potential for growth at a reasonable price. As a consulting firm to help soccer club managers find high-performing players in terms of value of players and profits for teams, we built an Ensemble model to find the GOLD players and to increase profits for each team.

Based on our initial intuition, as we mentioned in primary hypothesis, at the very beginning of the project our group thought that age, ball control, and acceleration would be significant to figure out excellent and profitable players to a club. In that sense, we built a Logistic Regression model using the three variables (i.e. Age, Ball control, and Acceleration) to predict who GOLD players are. However, the model gave us just 16% adjusted R- square and also Acceleration was not significant in terms of p-value (more than 0.05).

With 92.5% accuracy and 59.8% R-square for testing dataset using the Ensemble model that we built, we were able to predict who the GOLD players are more precisely. It means that we could figure out who more valuable players are for the club managers in terms of high value and more profits. Then, based on the model, we calculated new profits that each club could make from the model and were able to compare old profits from the original dataset. (we elaborated the computation logic behind the model in Profit/Loss Analysis section in detail.)

According to the graph below (Fig [14]), the average of old profits per player for all 70 clubs is -€5.59 million euros and the average of new profits per player for the clubs is -€5.02 million euros. So, using the Ensemble model that we built, the club managers can increase profits by €0.57 million euros per player in average. Then, the average number of players for each FIFA club is 25 so that the club manager for each club is able to increase profits by €14.31 million euros (€0.57 * 25 players = € 14.31) in general. Moreover, 32 (46%) out of 70 clubs gained more profits using our best model (i.e. the Ensemble model) by €2.28 million euros per players in average. In fact, the profits of most clubs are negative for both old and new, so it seems that they are losing money at all times. However, we just focused on improvements not either negative or positive values themselves because the difference between market value and the original value could be discrepant for some reasons such as time difference. Moreover, we could recognize that top clubs in terms of the difference between old and new are just tier 2 and tier 3. (Tier group is divided into 3 in terms of its performance in a league.) We assume the reason is that tier 1 clubs usually invest for top players and they don't have much potential to grow. However, tier 2 and tier 3 clubs invest money to players who have a high potential to grow in the future.

Initially we assumed that Age would be one of the most significant variables to predict GOLD players (the younger, the better), however, using stepwise regression and DOE analysis we were surprised Age is not significant from a statistical perspective. For that reason, we built our model without Age variable. According to the results, however, we found that 75% of our GOLD players (GOLD_Overall) are less

than 30 years old. At the same time, since the average of our predicted GOLD players is 27 years old, which is the peak of a soccer player, so possibly Age could have been an important distinguishing demographic feature.

Based on some additional insights which weren't quite noticeable before model building, our prescriptive analytics decisions would influence selection of players from five countries: Spain, Germany, Brazil, Italy, France. It made business sense as these countries have a lot of young rising talented players and few of the top players in FIFA. Fig[15], Fig[16]

K. Improvements

Improvements can be interpreted in many ways in terms of business point of view. As for this case, our group would focus more on improvements for better prediction of the GOLD players, if we start the project again from scratch. Here are two things to improve for the better results.

First of all, we would fully utilize domain expertise by consulting with experts such as a professional sports agency and a sports reporter for soccer. Since we do not have expertise for this area, we did not know what the objective for the project should be, what we need for a further analysis, and how to collect data we need at the beginning of the project. For this reason, we just learned from many trials and errors. If we consult and communicate with experts, we could spend less time and build a better model with some external data.

Second, we would enrich our dataset with some qualitative variables from external sources. Based on our dataset, we just have quantitative variables to predict Y output. If we have some qualitative data such as motivation, ability to adapt to new environments, and ability to speak multiple languages, we could build a model with these variables and get some better results. We believe that there are more factors to consider for us to predict Y output, and those qualitative variables can be significant to predict the GOLD players. In addition to this we could also perform feature engineering on a few variables that we believe could be an important predictor. For example, we could transform Age into 2 new features - one being Age < 27 and the other being Age >= 27.

L. Conclusion

To help FIFA clubs identify golden players, we imported information from external data, performed necessary transformations and feature engineering to create or select relevant expert input X variables. Then, we created a Logistic Regression Model and an Ensemble Model to predict the most profitable players more accurately. We are able to help the clubs acquire high-value players at a low cost, which can greatly reduce the clubs' loss or increase their profit (€0.57 million/player). We can also help the clubs form stronger teams that have a high chance of winning future games. We found that Ensemble Model works for more efficiently and accurately, both from a statistical and business perspective. Additionally, the important performance KPIs such as Positioning, Reactions, Composure, Ball control, Acceleration,

Free Kick Accuracy, can help us distinguish between players who are more likely to be profitable and those who are not.

M. Bibliography

I. Arif, A. (2018). Data Warehousing, Business Intelligence and Data Mining. [Presentation]. *Courses presented at Marshall School of Business*. University of Southern California, Los Angeles.

II. Kaggle.com. (2018). *FIFA 18 Complete Player Dataset*. [online] Available at: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset [Accessed 14 Nov. 2018].

III. Scikit-learn.org. (2018). *scikit-learn: machine learning in Python — scikit-learn 0.20.1 documentation*. [online] Available at: https://scikit-learn.org/stable/ [Accessed 20 Nov. 2018].

IV. Transfermarkt.com. (2018). *Most valuable clubs (Detailed view)*. [online] Available at: https://www.transfermarkt.com/vereins-statistik/wertvollstemannschaften/marktwertetop?plus=1 [Accessed 20 Nov. 2018].

N. Appendix

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
✓	✓	Intercept[0]	52.8407866	1	0	1
		Age	0	1	1.963652	0.16112
	✓	Wage(¬Ç in K)	-0.0272539	1	67.13827	2.5e-16
		Special	0	1	7.854601	0.00507
	✓	Acceleration	-0.0630503	1	11.8653	0.00057
		Aggression	0	1	0.519217	0.47118
		Agility	0	1	0.043143	0.83546
	✓	Balance	0.04190192	1	13.61045	0.00022
	✓	Ball control	-0.093813	1	2.686559	0.1012
	✓	Composure	-0.0841434	1	20.53665	5.85e-6
		Crossing	0	1	3.717781	0.05384
		Curve	0	1	0.255134	0.61348
		Dribbling	0	1	4.714426	0.02991
	✓	Finishing	-0.0538014	1	4.183666	0.04082
		Free kick accuracy	0	1	1.183778	0.27659
		GK diving	0	1	0.910464	0.33999
		GK handling	0	1	0.831226	0.36192
		GK kicking	0	1	0.096645	0.75589
		GK positioning	0	1	2.887e-7	0.99957
	✓	GK reflexes	-0.1435951	1	10.85142	0.00099
	✓	Heading accuracy	-0.0261848	1	5.814096	0.0159
		Interceptions	0	1	0.036774	0.84793
		Jumping	0	1	2.933786	0.08674
	✓	Long passing	-0.0534129	1	2.862861	0.09065
		Long shots	0	1	3.716699	0.05387
		Marking	0	1	4.318718	0.0377
		Penalties	0	1	2.393289	0.12186
	✓	Positioning	0.0952708	1	44.97124	2e-11
	✓	Reactions	-0.4076509	1	151.6839	7.4e-35

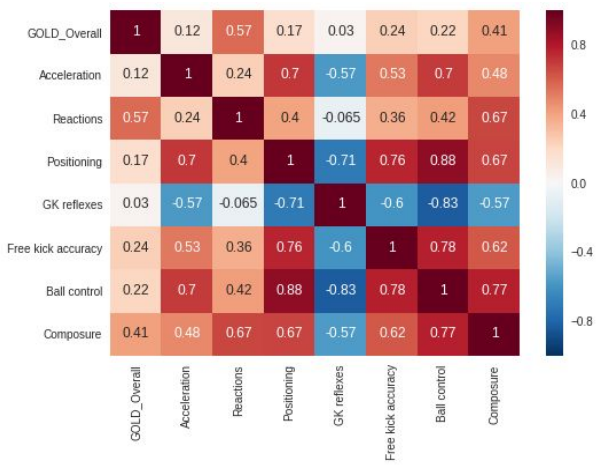
Fig[1]: Stepwise Regression

Screening for gold

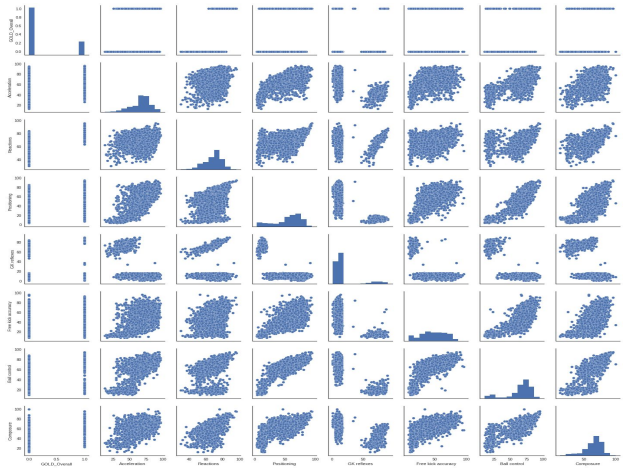
Contrasts

Term	Contrast		Lenth t-Ratio	Individual p-Value	Simultaneous p-Value
Reactions	0.199334		42.66	<.0001*	<.0001*
GK reflexes	0.027079		5.80	<.0001*	0.0002*
Composure	0.051369		10.99	<.0001*	<.0001*
Finishing	0.016338		3.50	0.0005*	0.7196
Positioning	-0.032977		-7.06	<.0001*	<.0001*
Ball control	0.019572		4.19	<.0001*	0.0800
Acceleration	0.013282		2.84	0.0051*	1.0000
Long passing	0.009263		1.98	0.0490*	1.0000
Balance	-0.007106		-1.52	0.1295	1.0000
Heading accuracy	0.002719		0.58	0.5602	1.0000
Reactions*Reactions	0.155034	*	33.18	<.0001*	<.0001*
Reactions*GK reflexes	0.028904	*	6.19	<.0001*	0.0001*
GK reflexes*GK reflexes	0.017575	*	3.76	0.0003*	0.3780
Reactions*Composure	0.060988	*	13.05	<.0001*	<.0001*
GK reflexes*Composure	-0.018581	*	-3.98	0.0001*	0.1828
Composure*Composure	0.014944	*	3.20	0.0011*	0.9637
Reactions*Finishing	0.008056	*	1.72	0.0851	1.0000
GK reflexes*Finishing	-0.000778	*	-0.17	0.8644	1.0000
Composure*Finishing	0.009562	*	2.05	0.0421*	1.0000
Finishing*Finishing	0.013556	*	2.90	0.0039*	0.9999
Reactions*Positioning	-0.021637	*	-4.63	<.0001*	0.0120*
GK reflexes*Positioning	-0.001830	*	-0.39	0.6916	1.0000
Composure*Positioning	-0.000069	*	-0.01	0.9889	1.0000
Finishing*Positioning	0.005297	*	1.13	0.2596	1.0000
Positioning*Positioning	0.003762	*	0.81	0.4215	1.0000
Reactions*Ball control	0.025200	*	5.39	<.0001*	0.0007*
GK reflexes*Ball control	-0.017806	*	-3.81	0.0002*	0.3224
Composure*Ball control	0.009084	*	1.94	0.0521	1.0000
Finishing*Ball control	0.020697	*	4.43	<.0001*	0.0288*

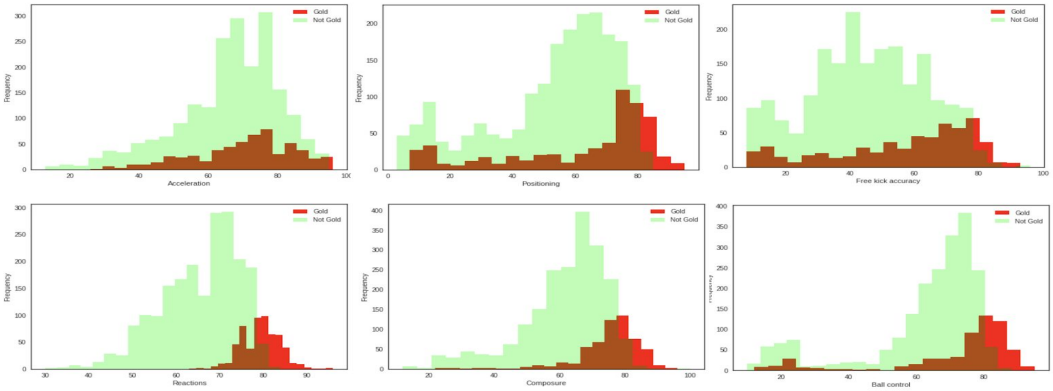
Fig[2]: DOE Screening Variables for Gold Players



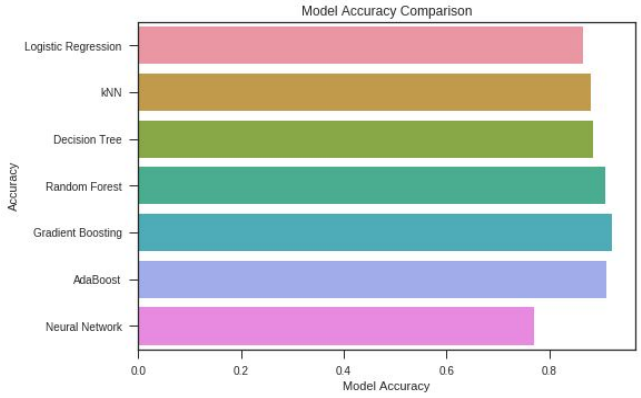
Fig[3]: Correlation Plot - Correlation of Variables



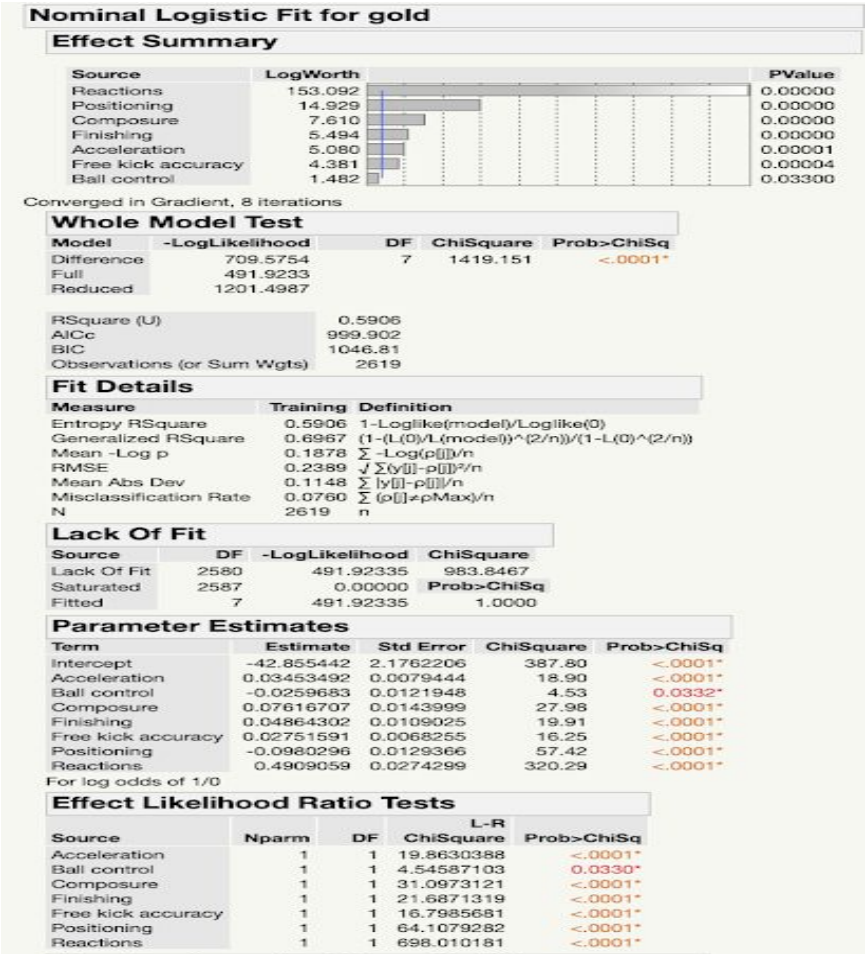
Fig[4]: Pair Plot - Correlation of Variables



Fig[5]: Histogram Plot - Important Variables Distribution



Fig[6]: Model Accuracy Comparison



Nominal Logistic Fit for gold

Effect Summary

Source	LogWorth	PValue
Reactions	153.092	0.00000
Positioning	14.929	0.00000
Composure	7.610	0.00000
Finishing	5.494	0.00000
Acceleration	5.080	0.00001
Free kick accuracy	4.381	0.00004
Ball control	1.482	0.03300

Converged in Gradient, 8 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	709.5754	7	1419.151	<.0001*
Full	491.9233			
Reduced	1201.4987			

RSquare (U)	0.5906
AICc	999.902
BIC	1046.81
Observations (or Sum Wgts)	2619

Fit Details

Measure	Training	Definition
Entropy RSquare	0.5906	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6967	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1878	$\sum$ -Log(p[j])/n
RMSE	0.2389	$\sqrt{\sum(y[j]-p[j])^2/n}$
Mean Abs Dev	0.1148	$\sum$ |y[j]-p[j]|/n
Misclassification Rate	0.0760	$\sum$ (p[j]≠pMax)/n
N	2619	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	2580	491.92335	983.8467
Saturated	2587	0.00000	Prob>ChiSq
Fitted	7	491.92335	1.0000

Parameter Estimates

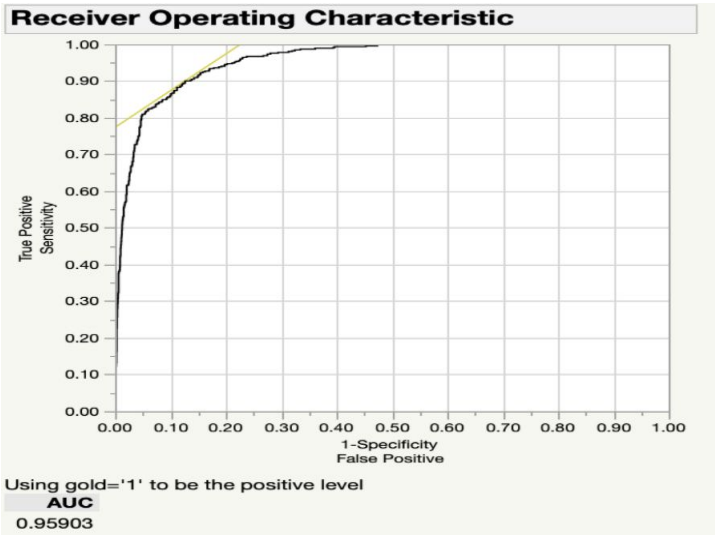
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-42.855442	2.1762206	387.80	<.0001*
Acceleration	0.03453492	0.0079444	18.90	<.0001*
Ball control	-0.0259683	0.0121948	4.53	0.0332*
Composure	0.07616707	0.0143999	27.98	<.0001*
Finishing	0.04864302	0.0109025	19.91	<.0001*
Free kick accuracy	0.02751591	0.0068255	16.25	<.0001*
Positioning	-0.0980296	0.0129366	57.42	<.0001*
Reactions	0.4909059	0.0274299	320.29	<.0001*

For log odds of 1/0

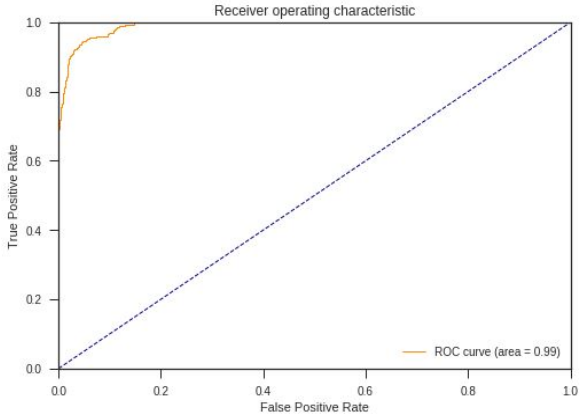
Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Acceleration	1	1	19.8630388	<.0001*
Ball control	1	1	4.54587103	0.0330*
Composure	1	1	31.0973121	<.0001*
Finishing	1	1	21.6871319	<.0001*
Free kick accuracy	1	1	16.7985681	<.0001*
Positioning	1	1	64.1079282	<.0001*
Reactions	1	1	698.010181	<.0001*

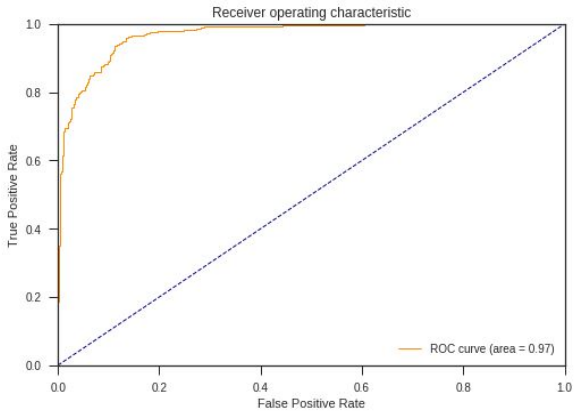
Fig[7]: Logistic Regression-training



Fig[8]: ROC curve for logistic regression-training



Fig[9]: ROC curve for Ensemble Model Training



Fig[10]: ROC Curve for Ensemble Model Testing

Nominal Logistic Fit for GOLD_Overall

Effect Summary

Source	LogWorth	PValue
Reactions	98.830	0.00000
Positioning	12.686	0.00000
Free kick accuracy	6.726	0.00000
Finishing	4.486	0.00003
Acceleration	3.118	0.00076
Composure	2.892	0.00128
Ball control	0.375	0.42160

Converged in Gradient, 8 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	421.77878	7	843.5576	<.0001*
Full	261.58529			
Reduced	683.36407			

RSquare (U)	0.6172
AICc	539.281
BIC	580.587
Observations (or Sum Wgts)	1309

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	1291	261.58529	523.1706
Saturated	1298	0.00000	Prob>ChiSq
Fitted	7	261.58529	1.0000

Parameter Estimates

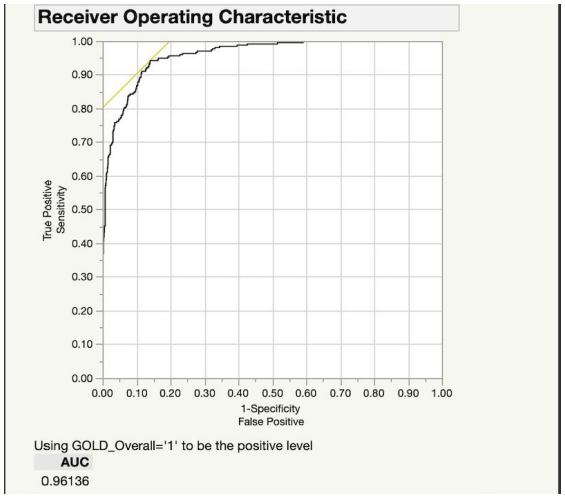
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-45.685005	3.2395676	198.87	<.0001*
Acceleration	0.03447109	0.0104261	10.93	0.0009*
Ball control	-0.0131993	0.0164036	0.65	0.4210
Finishing	0.05780239	0.0145567	15.77	<.0001*
Composure	0.05570927	0.0177079	9.90	0.0017*
Positioning	-0.1239159	0.0181136	46.80	<.0001*
Reactions	0.54409469	0.0408491	177.41	<.0001*
Free kick accuracy	0.04744309	0.0095116	24.88	<.0001*

For log odds of 1/0

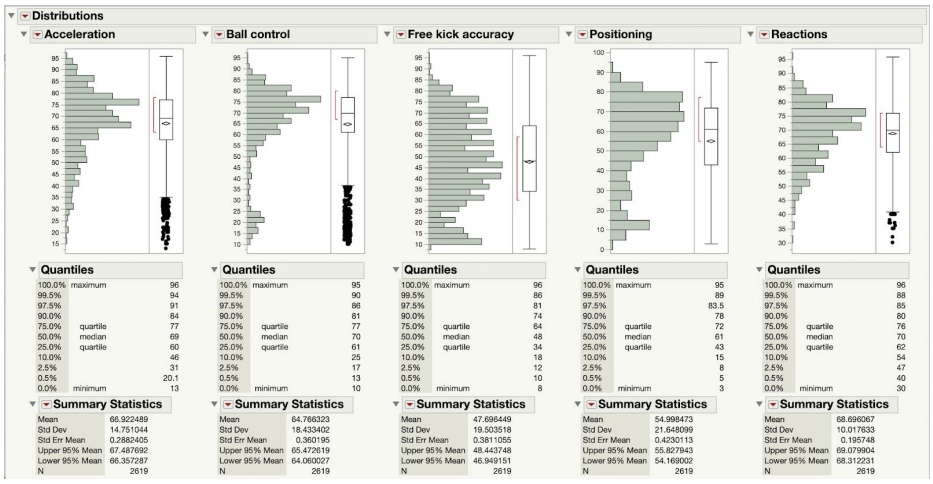
Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Acceleration	1	1	11.3308927	0.0008*
Ball control	1	1	0.6458372	0.4216
Finishing	1	1	17.2583695	<.0001*
Composure	1	1	10.3686418	0.0013*
Positioning	1	1	53.9470453	<.0001*
Reactions	1	1	448.566813	<.0001*
Free kick accuracy	1	1	27.1520275	<.0001*

Fig[11]: Logistic regression-testing



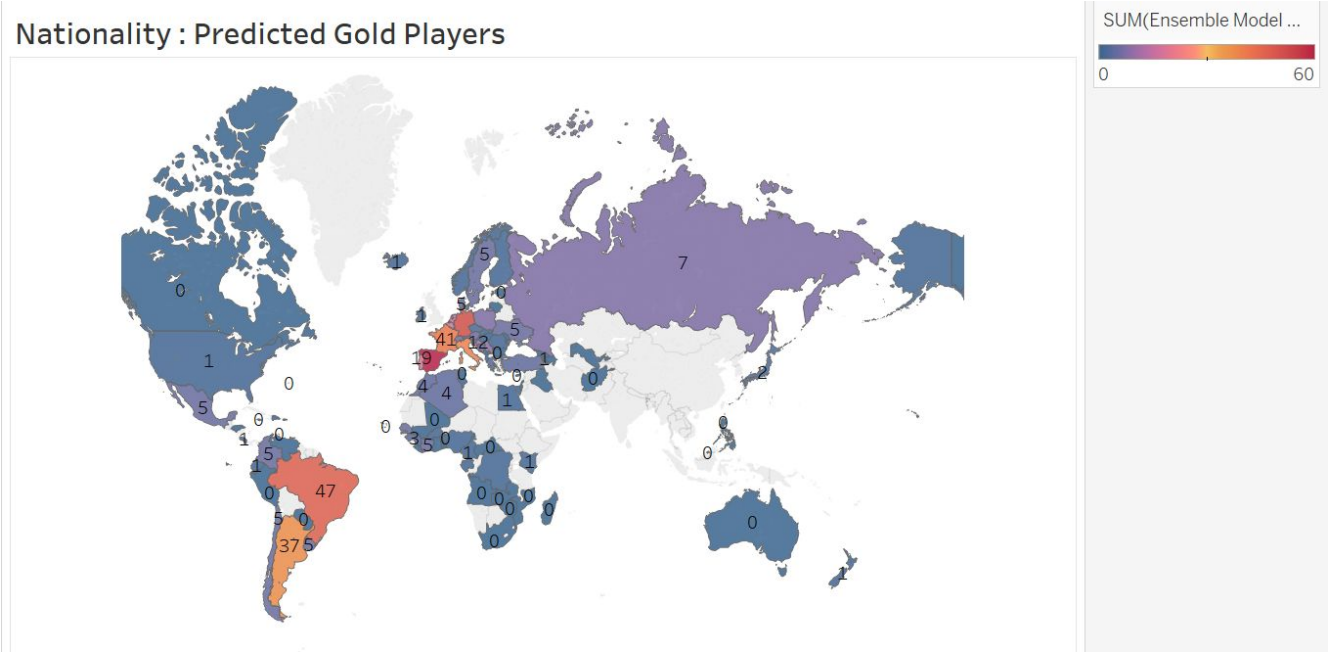
Fig[12]: Logistic Regression testing ROC curve



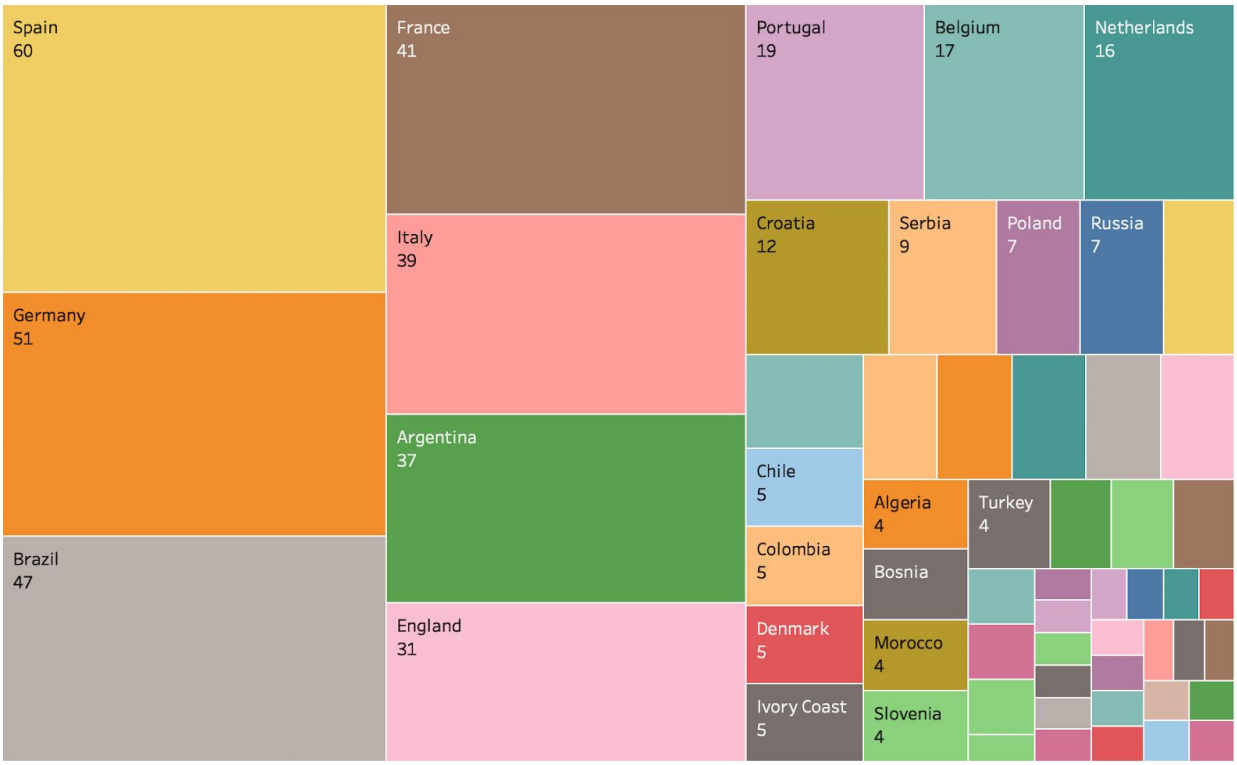
Fig[13]: Distribution of Variables

Club	Old	New	Difference
Average of Total	**-€ 5.59**	**-€ 5.02**	**€ 0.57**
Crystal Palace	-€ 9.70	-€ 0.82	€ 8.88
Atalanta BC	-€ 16.72	-€ 8.97	€ 7.75
ACF Fiorentina	-€ 6.83	-€ 0.33	€ 6.50
Ajax Amsterdam	-€ 12.23	-€ 6.48	€ 5.75
Torino FC	-€ 10.64	-€ 5.47	€ 5.17
RB Leipzig	-€ 12.00	-€ 6.91	€ 5.09
Stoke City	-€ 11.18	-€ 6.53	€ 4.65
AFC Bournemouth	-€ 2.99	€ 0.48	€ 3.47
Eintracht Frankfurt	-€ 8.25	-€ 4.94	€ 3.31
Burnley FC	-€ 3.43	-€ 0.18	€ 3.25
Real Betis Balompi V D	-€ 6.12	-€ 3.45	€ 2.67
VfB Stuttgart	-€ 3.39	-€ 0.79	€ 2.60
Leicester City	-€ 8.80	-€ 7.00	€ 1.80
FC Augsburg	-€ 3.22	-€ 1.47	€ 1.75
Hertha BSC	-€ 8.03	-€ 6.53	€ 1.50
Watford FC	-€ 5.27	-€ 3.99	€ 1.28
FC Porto	-€ 7.64	-€ 6.78	€ 0.86
West Bromwich Albion	-€ 5.87	-€ 5.04	€ 0.83
Sporting CP	-€ 13.35	-€ 12.54	€ 0.81
Borussia Dortmund	-€ 8.00	-€ 7.33	€ 0.68
Celta de Vigo	-€ 8.18	-€ 7.52	€ 0.66
Athletic Bilbao	-€ 9.64	-€ 9.02	€ 0.62
OGC Nice	-€ 12.55	-€ 12.01	€ 0.54
CSKA Moscow	-€ 6.74	-€ 6.24	€ 0.50
Zenit St. Petersburg	-€ 7.62	-€ 7.19	€ 0.44
FC Schalke 04	-€ 7.85	-€ 7.54	€ 0.30
VfL Wolfsburg	-€ 11.46	-€ 11.16	€ 0.30
AS Roma	-€ 5.37	-€ 5.10	€ 0.27
Hannover 96	-€ 8.93	-€ 8.68	€ 0.25
Spartak Moscow	-€ 11.73	-€ 11.50	€ 0.23
Borussia MV nchengladbach	-€ 8.87	-€ 8.65	€ 0.22
Villarreal CF	-€ 12.10	-€ 11.95	€ 0.15
AC Milan	-€ 6.15	-€ 6.15	€ 0.00
FC Girondins Bordeaux	-€ 15.44	-€ 15.44	€ 0.00
Getafe CF	-€ 5.30	-€ 5.30	€ 0.00
Inter Milan	-€ 1.72	-€ 1.72	€ 0.00
Juventus FC	€ 2.68	€ 2.68	€ 0.00
LOSC Lille	-€ 0.64	-€ 0.64	€ 0.00
Manchester City	€ 9.32	€ 9.32	€ 0.00
Paris Saint-Germain	-€ 2.01	-€ 2.01	€ 0.00
PSV Eindhoven	-€ 4.73	-€ 4.73	€ 0.00
Shakhtar Donetsk	-€ 10.98	-€ 10.98	€ 0.00
SS Lazio	-€ 6.39	-€ 6.39	€ 0.00
SV Werder Bremen	-€ 12.94	-€ 12.94	€ 0.00
UC Sampdoria	-€ 7.15	-€ 7.15	€ 0.00
Udinese Calcio	-€ 1.02	-€ 1.02	€ 0.00
US Sassuolo	-€ 13.93	-€ 13.93	€ 0.00
Real Madrid CF	-€ 2.31	-€ 2.37	-€ 0.06
Olympique Lyon	-€ 5.07	-€ 5.27	-€ 0.20
SL Benfica	-€ 2.88	-€ 3.23	-€ 0.36
Galatasaray SK	-€ 7.69	-€ 8.23	-€ 0.54
Bayer 04 Leverkusen	-€ 6.95	-€ 7.55	-€ 0.60
Sevilla FC	-€ 4.84	-€ 5.62	-€ 0.78
Valencia CF	€ 3.12	€ 2.31	-€ 0.81
TSG 1899 Hoffenheim	-€ 8.35	-€ 9.28	-€ 0.93
Liverpool	€ 11.55	€ 10.52	-€ 1.03
SSC Napoli	-€ 2.68	-€ 3.73	-€ 1.05
West Ham United	-€ 3.03	-€ 4.15	-€ 1.12
AS Monaco	-€ 16.39	-€ 17.60	-€ 1.21
Everton FC	-€ 3.06	-€ 4.44	-€ 1.38
Bayern Munich	-€ 5.57	-€ 7.17	-€ 1.60
Tottenham Hotspur	€ 1.44	-€ 0.17	-€ 1.62
FC Barcelona	€ 20.36	€ 18.73	-€ 1.64
Manchester United	€ 5.82	€ 4.17	-€ 1.65
Olympique Marseille	-€ 7.63	-€ 9.55	-€ 1.92
Real Sociedad	-€ 9.15	-€ 11.15	-€ 2.00
Arsenal FC	-€ 1.49	-€ 4.17	-€ 2.69
Lokomotiv Moscow	-€ 6.63	-€ 9.75	-€ 3.13
Atl√©tico Madrid	€ 16.95	€ 13.81	-€ 3.15
Southampton FC	-€ 5.65	-€ 9.19	-€ 3.55

Fig[14]: Profit/Loss Analysis with clubs with highest increase in profits at the top



Fig[15]: Predicted Gold Players 1 by Nationality



Fig[16]: Predicted Gold Players 2 by Nationality