# IDS575: BUSINESS ANALYTICS STATISTICS

**PROJECT – REPORT 1**

**Srujan Kadanur Viswanth (UIN: 659632968)**

**Nitin Philip (UIN: 660088429)**

**Project 1 – Transfer value prediction**

**Purpose:**

Football is a beautiful game. Football players earn a lot of money. Each player has a market value. The purpose of this project is to look into the various factors which can help determine the market value of football player.

**Data Summary:**

The dataset contains data on 461 football (soccer) players collected from the Fantasy Premier League, Wikipedia and transfermrkt.com. It has details on their playing statistics, their popularity and their market value.

The dataset has 17 variables. A description of each is given below.

Variable Details:

|  |  |  |
| --- | --- | --- |
| Variable | Description | Data Type |
| name | Name of the player | String |
| club | Club of the player | String |
| age | Age of the player | Numeric |
| position | The usual position on the pitch | String |
| position\_cat | Category of the position  - 1 for attackers  - 2 for midfielders  - 3 for defenders  - 4 for goalkeepers | Numeric |
| market\_value | As on transfermrkt.com on July 20th, 2017 | Numeric |
| page\_views | Average daily Wikipedia page views from September 1, 2016 to May 1, 2017 | Numeric |
| fpl\_value | Value in Fantasy Premier League as on July 20th, 2017 | Numeric |
| fpl\_sel | % of FPL players who have selected that player in their team | String |
| fpl\_points | FPL points accumulated over the previous season | Numeric |
| region | Region player belongs to  - 1 for England  - 2 for EU  - 3 for Americas  - 4 for Rest of World | Numeric |
| nationality | Citizenship of the player | String |
| new\_foreign | Whether a new signing from a different league, for 2017/18 (till 20th July) | Numeric |
| age\_cat | Age category | Numeric |
| club\_id | Club ID | Numeric |
| big\_club | Whether one of the Top 6 clubs | Numeric |
| new\_signing | Whether a new signing for 2017/18 (till 20th July) | Numeric |

This dataset was obtained from the link: <https://www.kaggle.com/mauryashubham/english-premier-league-players-dataset>

**Hypothesis:**

In this dataset, the target variable will be the market value of the player. We have made the below hypothesis.

Hypothesis 1:

Expected variables like players age and performance (measured by FPL points) effect the transfer value.

Hypothesis 2:

Players popularity (measured by page views on Wikipedia) has a positive effect on the transfer value.

**Methods:**

1. We will first clean the data (if required) and then assess all the variables to find relationships between predictors.

2. The primary goal of this model is to build a linear regression model to predict the market value of a player. Certain players may be omitted if they don’t have a market value assigned.

3. Based on the regression model we can explain a player’s transfer value, estimate the impact of English Premier League’s popularity of the players who have been signed from various other leagues. Do a comparative analysis of the top 6 teams versus the remaining teams

4.For this we need to understand how the market values are distributed and does the market value of a player change if he is being signed by the top 6 team and also if the same player is being signed by a club outside top 6.

5. At the end we find out if there is any correlation between the market value and popularity of the player and if the age of a player plays a significant role in determining the market value. Building a decent model by diving deep into the data is the main focus of this project.

**Project 2 – Real Time Bidding (Ambitious)**

**Purpose:**

This is real real-time bidding data that is used to predict if an advertiser should bid for a marketing slot e.g. a banner on a webpage. Explanatory variables are things like browser, operation system or time of the day the user is online, marketplace his identifiers were traded on earlier, etc. The column 'convert' is 1, when the person clicked on the ad, and 0 if this is not the case.

**Data Summary:**

The dataset contains 1 million records with 88 principle components that can be used to predict our target variable convert.

The dataset has 88 variables. A description of them is given below.

Variable Details:

|  |  |  |
| --- | --- | --- |
| Variable | Description | Data Type |
| 0 | Principle Component 0 | Numeric |
| 1 | Principle Component 1 | Numeric |
| 2 | Principle Component 2 | Numeric |
| 3 | Principle Component 3 | Numeric |
| 4 | Principle Component 4 | Numeric |
| 5 | Principle Component 5 | Numeric |
| . | . | . |
| . | . | . |
| . | . | . |
| 85 | Principle Component 85 | Numeric |
| 86 | Principle Component 86 | Numeric |
| 87 | Principle Component 87 | Numeric |
| convert | ads that get clicked 1; ads that are only shown 0 | Numeric |

This dataset was obtained from the link: <https://www.kaggle.com/zurfer/rtb>

**Hypothesis:**

In this dataset, the target variable will be convert. We have made the below hypothesis.

Hypothesis 1:

As the data is heavily imbalanced, we would not be training for accuracy, but rather try for a good AUC.

Hypothesis 2:

On comparing models like logistic regression, decision trees and SVM, we can find the best fit.

**Methods:**

1. We will first clean the data (if required) and then assess all the variables to find relationships between predictors.

2. As the data is heavily imbalanced, we will under sample the data and obtain both classes 1:1.

3. We will then create models using algorithms for Logistic Regression, Random Forest and SVM.

4. Next would be to compute and compare different evaluation metrics to see if the models are performing well on the test data.