Data Programming

Project 1

"Statement of Academic Honesty:

The following code represents our own work. We have neither received nor given inappropriate assistance. We have not copied or modified code from any source other than the course webpage or the course textbook. We recognize that any unauthorized assistance or plagiarism will be handled in accordance with Georgia State University's Academic Honesty Policy and the policies of this course. We recognize that our work is based on an assignment created by the Institute for Insight at Georgia State University. Any publishing or posting of source code for this project is strictly prohibited unless you have written consent from the Institute for Insight at Georgia State University."

Team members:

Arpil Mehta	*amehta21@student.gsu.edu*
Bhoomika Jaggi	*bjaggi1@student.gsu.edu*
Jagdish Pusuluru	*jpusuluru1@student.gsu.edu*
Piyush Godbole	*pgodbole2@student.gsu.edu*
Sagar Mehta (Team leader)	*smehta18@student.gsu.edu*

1. ***Show overall descriptive statistics of your dataset, number of data points, number of descriptive features, type of features, your target feature, and its type.***

Answer: The number of datapoints we have in our dataset are (461, 17) and the descriptive features are as follows:

- age: Age of the player (Type: int)
- fplvalue: Value in Fantasy Premier League as on July 20th ,2017 (Type: float)
- big_club: if the player is from top 6 clubs. (Type: int)
- page_views: Average daily Wikipedia page views from Sep 1, 2016, to May 1, 2017. (Type: int)
- position_cat: Assigned 1 for attackers, 2 for midfielders, 3 defenders, 4 for goalkeepers. (Type: int)

The target feature used in our project is

- market_value: It is the market value of the player as on transfermrkt.com on July 20th, 2017. (Type: float)

	count	mean	std	min	25%	50%	75%	max
age	461.0	26.804772	3.961892	17.00	24.0	27.0	30.0	38.0
position_cat	461.0	2.180043	1.000061	1.00	1.0	2.0	3.0	4.0
market_value	461.0	11.012039	12.257403	0.05	3.0	7.0	15.0	75.0
page_views	461.0	763.776573	931.805757	3.00	220.0	460.0	896.0	7664.0
fpl_value	461.0	5.447939	1.346695	4.00	4.5	5.0	5.5	12.5
fpl_points	461.0	57.314534	53.113811	0.00	5.0	51.0	94.0	264.0
region	460.0	1.993478	0.957689	1.00	1.0	2.0	2.0	4.0
new_foreign	461.0	0.034707	0.183236	0.00	0.0	0.0	0.0	1.0
age_cat	461.0	3.206074	1.279795	1.00	2.0	3.0	4.0	6.0
club_id	461.0	10.334056	5.726475	1.00	6.0	10.0	15.0	20.0
big_club	461.0	0.303688	0.460349	0.00	0.0	0.0	1.0	1.0
new_signing	461.0	0.145336	0.352822	0.00	0.0	0.0	0.0	1.0

2. ***Explore your features further in their distributions and plot their box plots. Show outliers for each feature. Do you think any of the outliers may impact your analysis? Why? Provide supporting visualizations with their analysis.***

Answer:

The fplvalue and page_views feature has several outliers which can't be removed since star players have outstanding fplvalue and more count of page_views as compared to regular players which plays an important role in predicting the market value of that player.
So, keeping these outliers is important as it will affect our analysis. For boxplots and other supporting visualizations, please refer to the code.

3. Determine if any features have missing data and what should be done with the missing data. Explain why the decision was made for each feature. If there is no missing data, explain how you handle missing data and why. Provide supporting visualizations with their analysis.

Answer:

Although, in this project, our dataset doesn't contain any missing values in the descriptive features we are using for our model. There are various possible ways to handle missing values such as:

1. *Calculate the average for that feature and fill the missing value with this average value.*
2. *Calculate the median for that feature and fill the missing value with this average value.*
3. *Fill in zeroes for the feature that have missing values.*
4. *Drop the row that have missing values (Not a recommended method to handle missing values)*

4. What data pre-processing do you apply? E.g., encoding features, missing values, scaling, etc. Explain each process and why you use it.

Answer:

The data is very well scraped and doesn't require much preprocessing. For training model 3 and model 4 we use one hot encoding and convert the *pos_cat* column to numeric by using *pd.get_dummies(). There is only one missing value in the region column, but we do not use the region column in our model, so we don't have to treat it.*

5. Analyze the balance or distribution of your target variable. Do you think any of these will present a problem and why? Provide supporting visualizations with their analysis.

Answer:

Based on the dataset available, our target variable is market_value which is right skewed. It seems obvious since we have a small pool of star players compared to regular players. (For visualization, please refer jupyter notebook)

6. What kind of ML approaches and algorithms do you take and why? E.g., supervised, regression, classification, binary, multi-class, split rate of data, logistic regression, SVM, decision trees etc. Provide supporting visualizations with their analysis.

Answer:

Considering the volume of the dataset, we used linear multiple regression model. We planned to predict the value of a variable based on the value of other variables. Here, the predicted variable market_value, is dependent on the features listed in question 1.

7. What evaluation metrics you used to evaluate the performance of your model. Discuss the results of your model as to which model performs better and why this would be the case. How would your model perform based on the results? What shortcomings your model has and possible implications?

Answer:

We will evaluate our model based on the mean squared error (MSE) for the model. We will select the model which has the least MSE. Based on our analysis, we conclude that the model no. 4 has performed comparatively better than other 3 models as it has the lowest training MSE. The training dataset on which our model is trained is not large since there is a cap of 25 players in each league team and we have data for almost 20 teams.

8. What is the solution you propose and implement to solve the problem? Explain your thought process as to what kind of stages and processes you have gone through to make decisions in each step. For instance, what led you to choose the evaluation metric you use? what motivated your selection of ML algorithms for prediction? why did you choose the preprocessing techniques you used? Provide supporting visualizations with their analysis?

Answer:

So, at the start of the project our goal was to predict market value of the player, in our case, since we had many features, the best way was to use linear regression algorithm, which would allow us to play with the features and give us the relevancy of each independent variable w.r.t to dependent variable. Linear Regression algorithm helps us best to solve the problem in hand and MSE helps to evaluate the model. Mean squared error (MSE), which is simply the averaged value of the MSE cost that we minimized to fit the linear regression model. The MSE is useful for comparing different regression models or for tuning their parameters via grid search and cross-validation, as it normalizes the MSE by the sample size.

Based on our analysis, we conclude that the model no. 4 has the least training MSE and will perform better than the above models, thus we can conclude that independent variables used in model 4 help us come close to the actual values.

(For visualization, please refer jupyter notebook)