**Identification of the target audience and their characteristics**

Pavel Hanekov

10 October, 2019

1. **Introduction**

Problem and interest: game-companies need to choose their target audience and want to understand how to make it right, they need to choose right parameters and compare them using ML to understand an impact on the result of each variable.

1. **Data acquisition and cleaning**

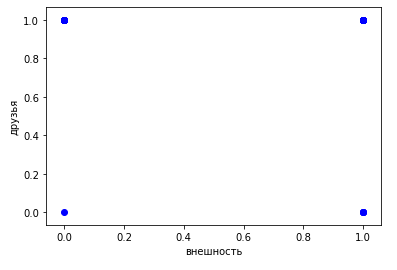
All data was collected by myself a year ago and proved with links from polls and open-data sources, it’s shown in yes/no answers, it doesn’t need cleaning. Feature selection consists of parameters(1/0 data variables) and sex and age.

1. **Exploratory Data Analysis**

It consists of showing all data on the map to make it easier understand the real situation



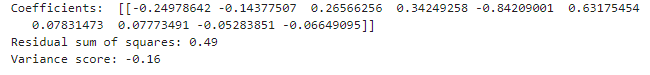
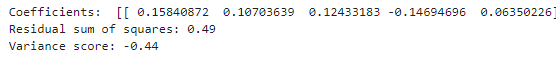
Scatter plot won’t help with analysis as there’re only 4 points



We can’t build a function using these points

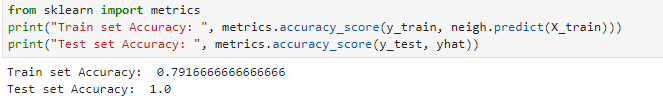
So it means that we need to choose another plot to make predictions on it

Linear regression shows low level of variance score which means lack of interconnection between data (different parameters were used)

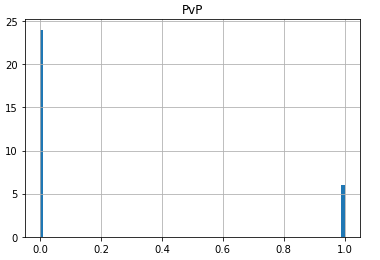


K-Nearest Neighbors technique

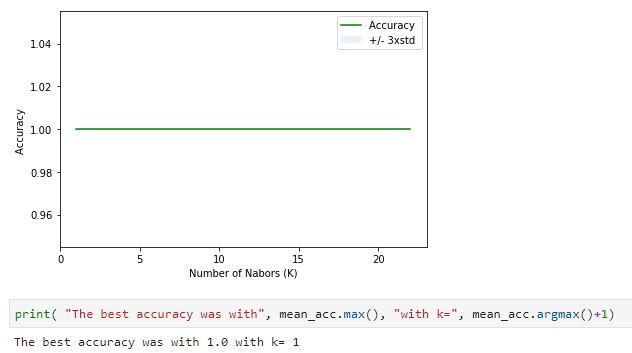
* shows high level of test set accuracy (1) and good 0.8 train set accuracy



* Histogram deals with 0/1 data better, showing everybody attitude to PvP (example)

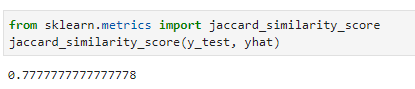


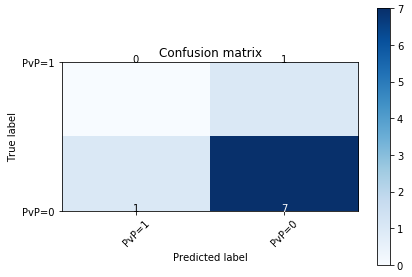
* Because of such a huge accuracy the right K=1



Logistic regression (all parameters are used)

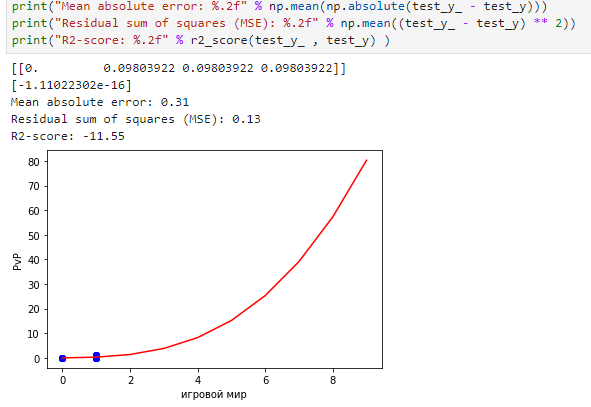
* Jaccard index is good



* Matrix illustrates right opinion to PvP(example), according to this we can conclude not to use such parameter as most people are ignoring it 

Polynomial regression

* Doesn’t have enough points to build a better plot and to predict the values in a better way
* Because of that huge R2-score



1. **Conclusions and future directions**

Matrix illustrated results the best way, but other tools were not accurate enough.

Maps are cool as illustrating data on the map is always a winning condition.

The problem was in using data which consists of 1/0 information, it's alright if there's only 1 column of 1/0 data and we are going to predict it, but not when all data is fulfilled with it, try to avoid it or add some extra columns.