**PUBG Analysis: Using Different Models to Predict Pro-Social Behavior**

“PlayerUnknown's Battlegrounds” (PUBG) is an [online](https://en.wikipedia.org/wiki/Online_game) [multiplayer](https://en.wikipedia.org/wiki/Multiplayer_video_game) [battle royale game](https://en.wikipedia.org/wiki/Battle_royale_game) developed and published by [PUBG Corporation](https://en.wikipedia.org/wiki/PUBG_Corporation), a [subsidiary](https://en.wikipedia.org/wiki/Subsidiary) of South Korean video game company [Bluehole](https://en.wikipedia.org/wiki/Bluehole_(company)" \o "Bluehole (company)).  In the game, up to one hundred players parachute onto an island and scavenge for weapons and equipment to kill others while avoiding getting killed themselves. The available safe area of the game's map decreases in size over time, directing surviving players into tighter areas to force encounters. The last player or team standing wins the round.

The PUBG game Data Service and API make it easier for users to have open access to in-game data, which developers, players and fans alike could use to conduct new and exciting researches or build interesting tools and services. In this study, I chose a random sample out of data collected on games played between the dates 19/11/2019 and 03/12/2019. The current sample consists out of approximately 56 thousand data-points which are described through 45 variables.

In this paper, I chose to predict the number of positive engagements between players in the game as quantified by the number of revives each player made to his fellow teammate. I chose to predict this number via multiple variables that are presented in the collected dataset. These variables reflect the actions behaviors and personality of the participant in a virtual gaming environment. Data that represents personal real-life information, is not presented here. This Data us both less-ethical to collect by a gaming company and is less relevant to my study.

Example of Several Variables:

“revives”: Number of times this player revived teammates.

“DBNOs”: Number of players knocked.

“assists”: Number of enemy players this player damaged that were killed by teammates.

“boosts”: Number of boost items used.

“damageDealt”: Total damage dealt. Note: Self-inflicted damage is subtracted.

“deathType”: The way by which this player died, or alive if they didn't.

“headshotKills”: Number of enemy players killed with headshots.

“heals”: Number of healing items used.

“killPlace”: This player's rank in the match based on kills.

“killStreaks”: Total number of kill streaks.

“kills”: Number of enemy players killed.

“longestKill”: longest Time in seconds between kills or until the first kill.

“name”: PUBG IGN of the player associated with this participant.

“playerId”: Account ID of the player associated with this participant.

“rideDistance”: Total distance traveled in vehicles measured in meters.

“roadKills”: Number of kills while in a vehicle.

“swimDistance”: Total distance traveled while swimming measured in meters.

“teamKills”: Number of times this player killed a teammate.

“timeSurvived”: Amount of time survived measured in seconds.

“vehicleDestroys”: Number of vehicles destroyed.

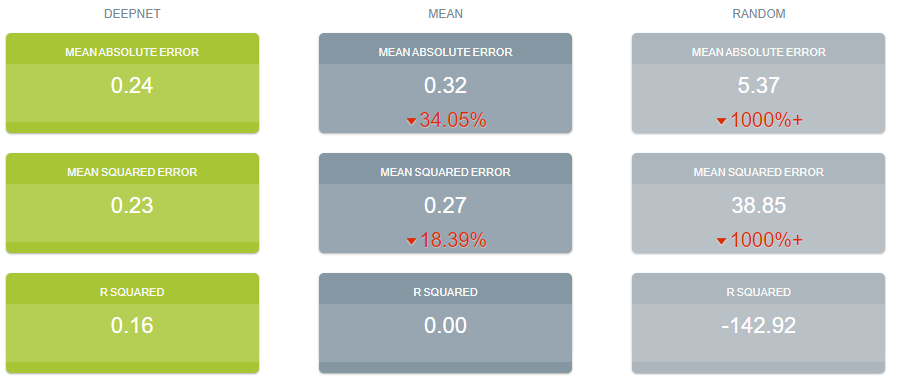
“walkDistance”: Total distance traveled on foot measured in meters.

“weaponsAcquired”: Number of weapons picked up.

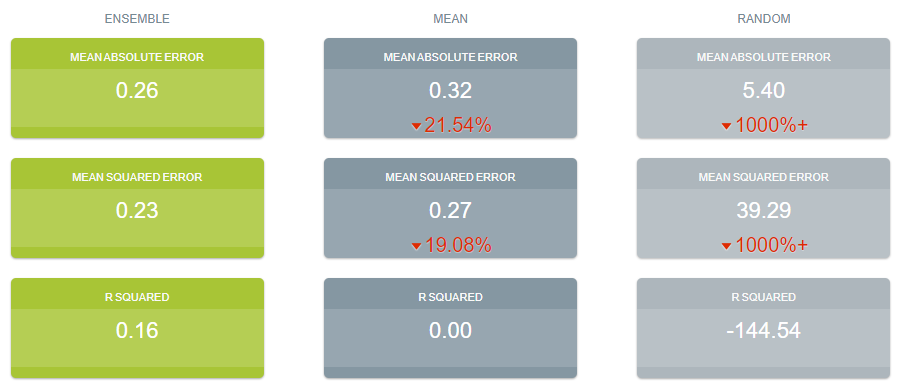
“winPlace”: This player's placement in the match.

The predicted variable, the number of positive engagements between players in the game as quantified by the number of revives each player made to his fellow teammate, was predicted using 4 different analysis methods. Linear Regression, neural networks, decision trees, and random forests. Each of these methods was conducted using 80% of the data (train set) and included automatic optimization with 5 hour-long optimization spans. Then, each model was evaluated by using the rest of the data (test data) which was 20% of the total data set. The success of this evaluation was quantified by the R2 value and the mean absolute error (MAE).

The results show that the neural networks model explained the largest amount of variance of the “revives” variable 16%, and produced the smallest number of the mean absolute error 24% (see fig. 1). Other models showed similar results: random forest- R2 =16%, MAE = 26% (see fig. 2), decision tree- R2 =15%, MAE = 27% (see fig. 3). The optimized linear regression models did not increase the explained variance of the “revives”: R2=-1.4817\*1011, MAE = 200,000 (see fig. 4).



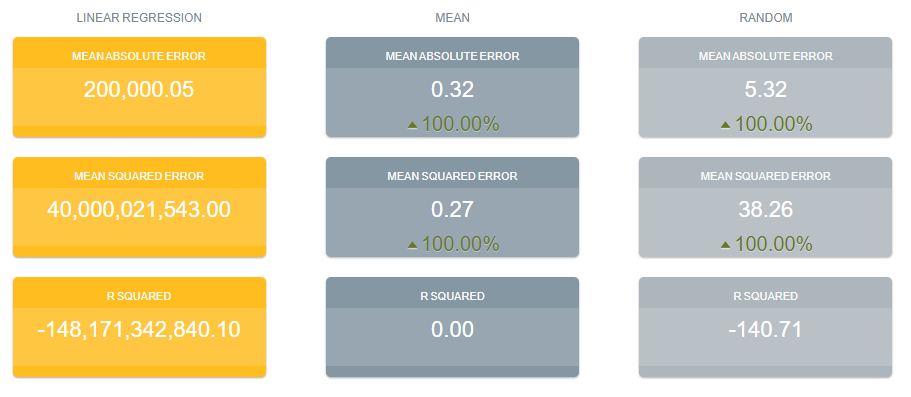
**Fig. 1**. Prediction of the variable “revives” using the analysis method of the neural networks model, as compared to prediction using mean or random values.



**Fig. 2**. Prediction of the variable “revives” using the analysis method of the random forest model, as compared to prediction using mean or random values.



**Fig. 3**. Prediction of the variable “revives” using the analysis method of the decision tree model, as compared to prediction using mean or random values.



**Fig. 3**. Prediction of the variable “revives” using the analysis method of the linear regression model, as compared to prediction using mean or random values.

**Bibliography:**

BigML, Inc. (2011). Retrieved from BigML: https://bigml.com

PUBG Corporation. (2019). *PUBG API Documentation*. Retrieved from PUBG: https://documentation.pubg.com/en/index.html