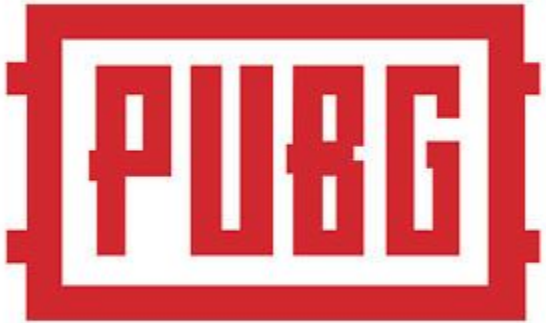


A data analysis on how to win first place in Players Unknown Battlegrounds

By Ameen Abdelghani

November 2018

What is the Data?

- Players Unknown Battlegrounds (PUBG) is the 5th best selling video game of all time
- Benchmark for battle royale genre
- Battle royale genre: 100 players whether in teams or solo spawn without any items, must scavenge for loot, the map is shrinking to a random location, last man or team standing wins

Business Problem

- PUBG developer's, Bluehole, since the initiation of this project, have not implemented any sort of strategy guide
- Until recently there wasn't even a training mode to practice
- Only one team / person can win a match, which a full match takes about 30 minutes, so the genre can get frustrating for players
- Players have differing opinions on the best way to win a match

Objective

- Determine the dependent variables to winning a match
- Does hiding and non-aggressive play style pay off?
- Does constantly moving, tracking, and being gung-ho work better?

VS

Why?

- Frustrated consumers of any industry will likely mean a drop off on usage of the product, which has happened to pubg in recent months
- Analyzing this data could help in constructing a strategy guide or simple tip guide to PUBG, creating more engaged players

The Process

- I'll be following a typical data science pipeline, which is called "OSEMN" (pronounced awesome).
- Obtaining the data
- Scrubbing or cleaning the data is the next step.
- Exploring the data will follow right after and allow further insight of what our dataset contains.
- Modeling the data will give us our predictive power
- Interpreting the data is last. With all the results and analysis of the data, what conclusion is made?

The Dataset

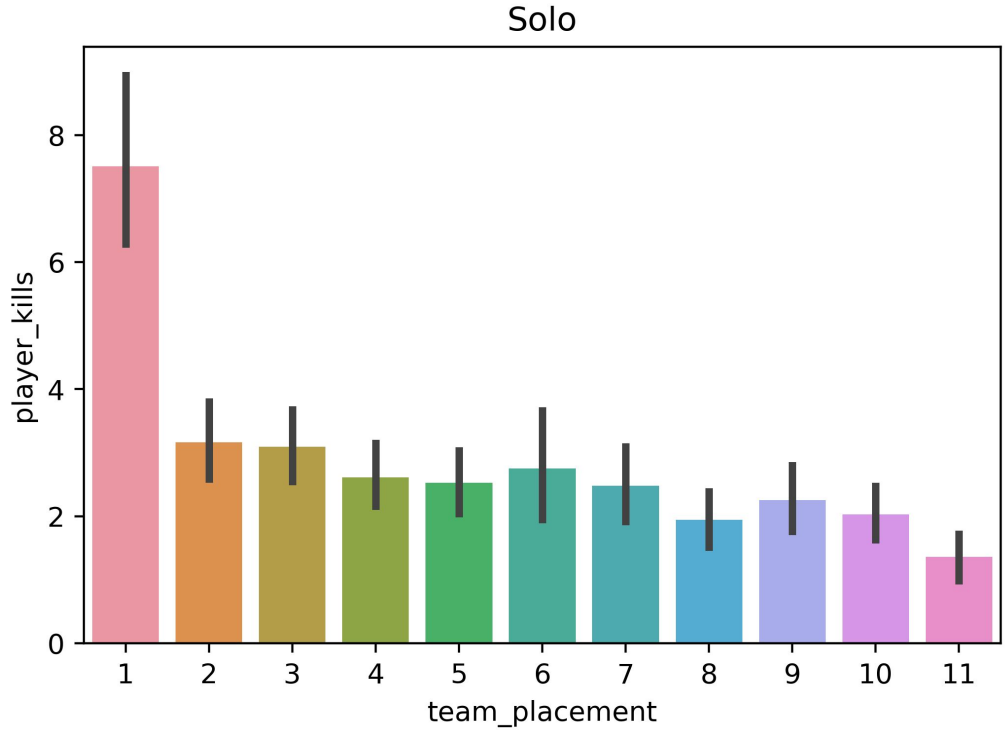
- The dataset can be found on kaggle.com
- No missing values
- Contained few outliers
- Columns include
  - Team placement
  - Damage
  - Kills
  - Knockdowns
  - Distance walked
  - Distance driven
  - Time survived
  - Party size

Cleaning

- Converted survival time from seconds into minutes
- Separated data for the three party types
    - Solo
    - Duo
    - Squad
- Converted party type into the category data type to save memory
- Removed bottom half of teams
- Removed observations that didn’t survive long, or barely walked

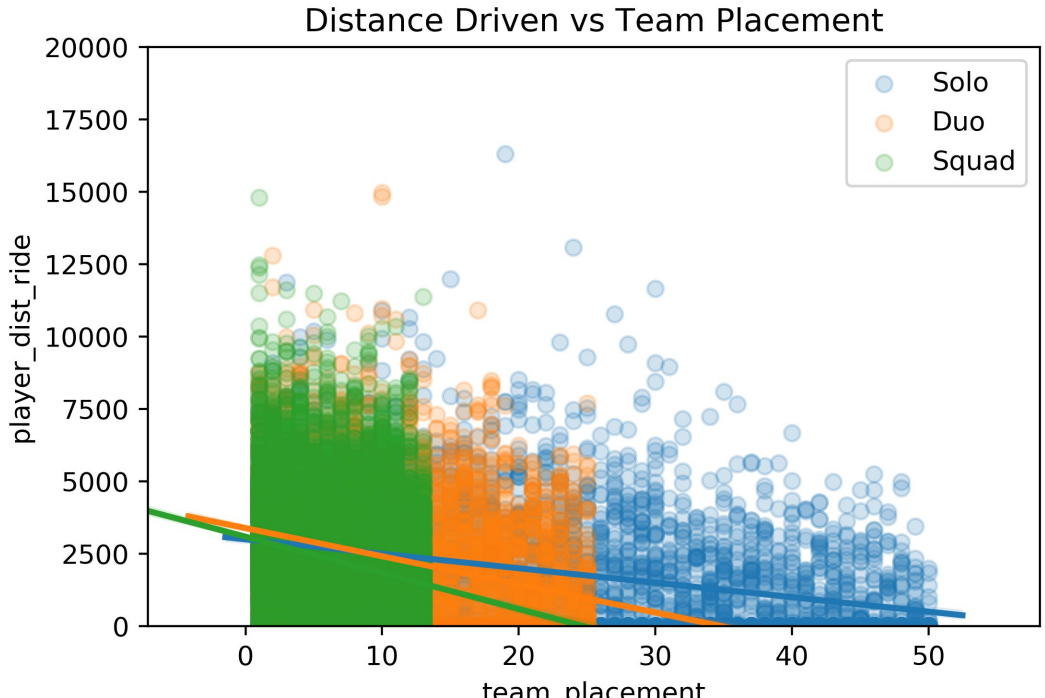
Data Exploration

Average kills of first place is 7, well above the remaining 10 placements

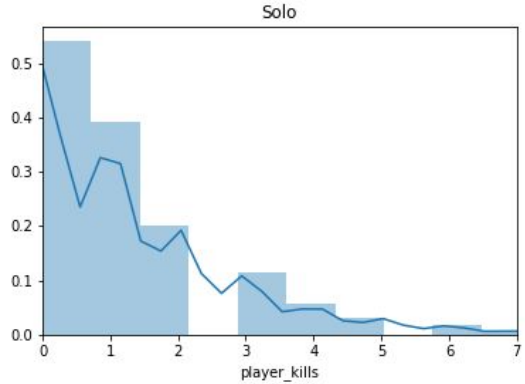


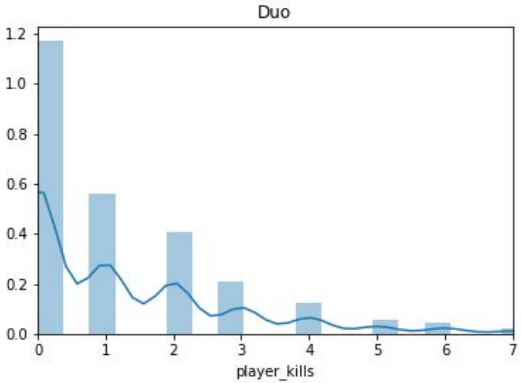
Data Exploration

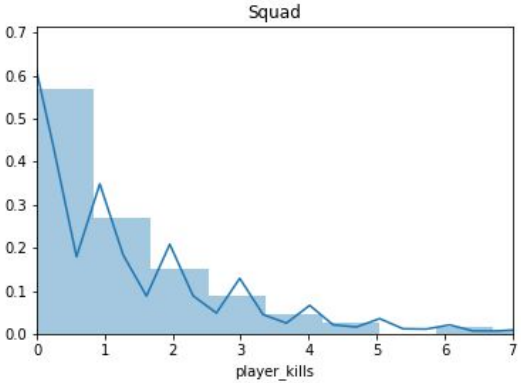
Driving a vehicle appears to work better for squad matches



Distribution of Kills

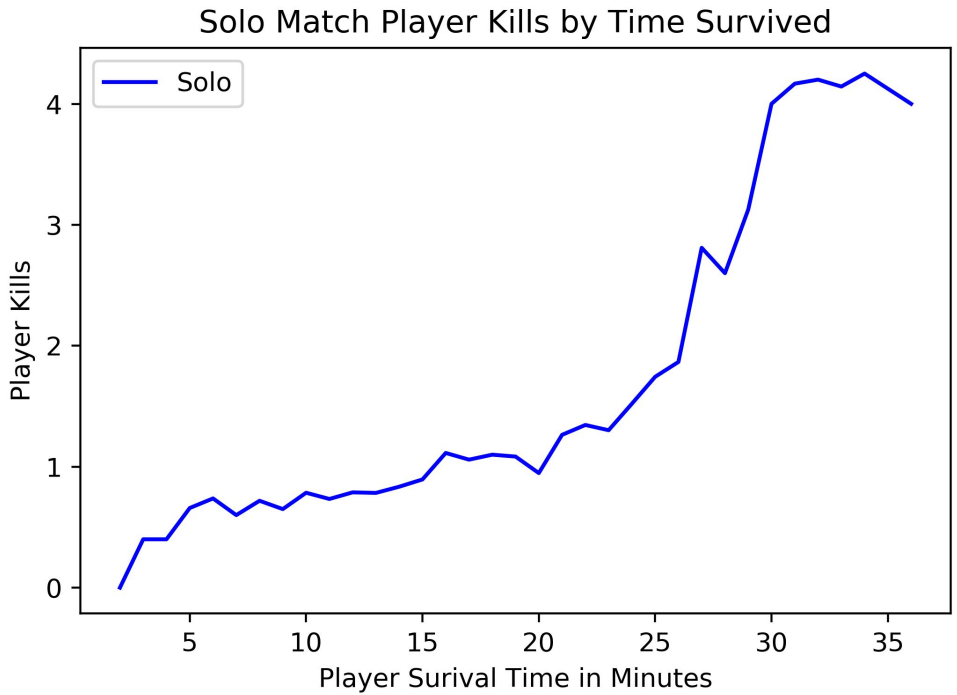






EDA

Mean survival time of players during each minute of survival



Statistical Analysis

- Pearson correlation with team placement against all other variables
- Kills, Damage, Distance Walked, and Distance driven highest correlations
- Negative correlation desired since better team placement is the lowest number
    - Kills : -.25 to -.39
    - Walking Distance: -.4 to -.55 (for solo only -.04)
    - Driving Distance : -.21 to -.26

Stats

- Kills vs Team Placement
  - T-Statistic = -92 P-value : 0. Indefinitely correlated
- Distance Driven vs Survival Time
  - Solo: P-value = 6.1.
  - Duo: P-value = 0
  - Squad: P-value = 0
- Interpretation: Driving a vehicle in solos is known to be very risky
- First Place in solo matches : Confidence Interval [6.2 - 8.9]

Building a Model

Objective: Building a classifier model to predict if a player will win first place given their stats in a match

1. Creating dummy variables by turning non first placements into 0
2. Drop irrelevant columns known by domain knowledge
3. Balance the dataset so 20% of the data first place
4. Compute the variance inflation factor of each column to rule out collinearity in the data
5. Compute the information value of each variable to determine predictive power of each variable
6. Train-split-test the data
7. Run Logistic Regression and Random forest classifier, and compare scoring of each model weighed by auc score and classification report

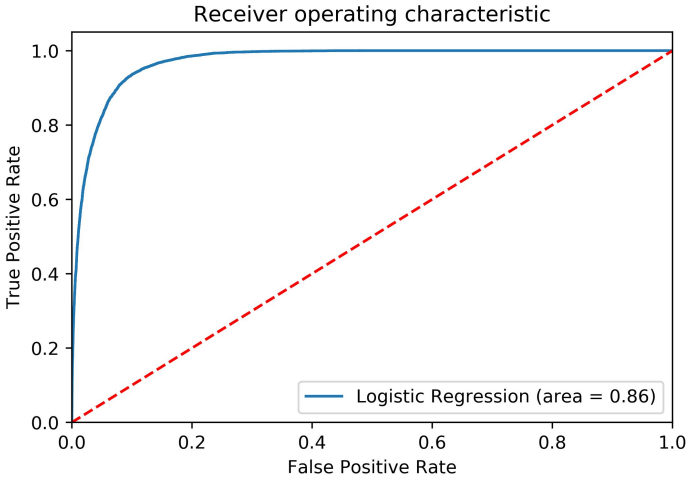
Feature Selection

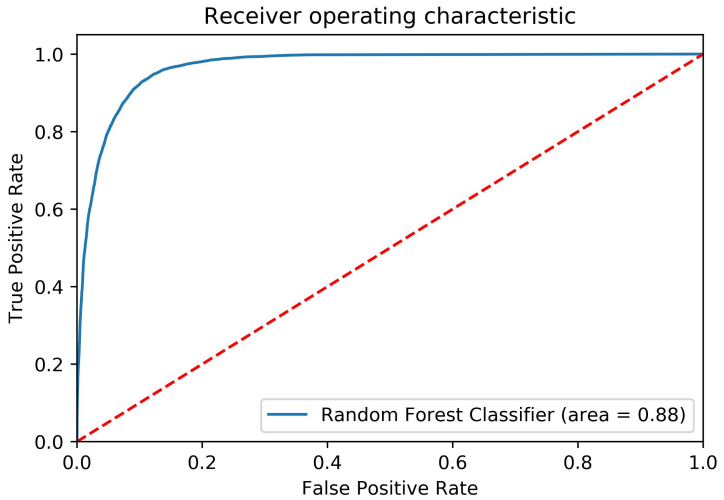
- Player Damage:
  - IV: 5.7
- Distance Walked
  - IV: 3.3
- Distance Driven
  - IV: 1.1

Winner: Random Forest Classifier

Area Under Curve Score

- Logistic Regression : 86%
- Random Forest Classifier: 88%





Recommendation

- How to win?
  - Don’t stay in one location, migrate!
  - Shoot to kill
  - Utilize vehicles
- Why?
  - Killing and looting bodies = Better equipment & Clearing areas
  - Moving around = Scouting environment, being the first to shoot
  - Utilizing Vehicles = Not getting zoned, first to loot, superior positioning

Future Endeavors

- Newer datasets have begun coming out which contain more features
- May induce greater accuracy on our model
- Location data can be used to determine fighting hotspots and where zone likely end
- Data on player accuracy might show new insights