Two pages project summary

More in-depth analysis can be found in the Jupyter notebook.

Data

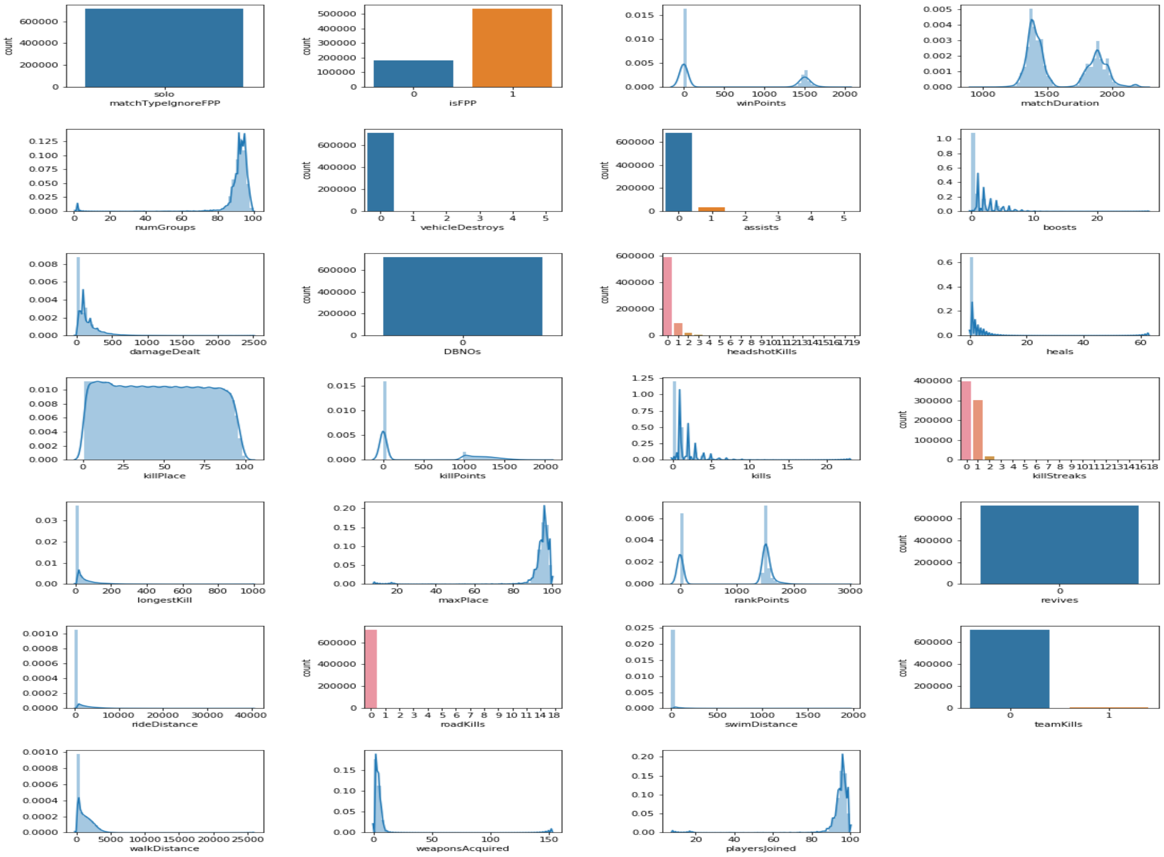
The data describes approximately 4mln matches of popular game PUBG (Players Unknown Battleground). It consists of 4 mln observations and 33 variables. Each observation represents a different game (basically every observation is different entry to the game). In this game only one player per match can win, i.e. there is 100 players entering each match, yet only one can survive! (It's basically more survival game than shooter). Data itself is a part of Competition Data on Kaggle, where winning prize is SWAG.

Data manipulation

Since the dataset consists mostly of continuous variables, we had to investigate them deeply to check their distribution, meanings and information value. We decided to change couple of them to binary variables, as values cumulated around 2-3 most frequent value which were really small, and frequency of its occurrence was quite rare. We also established division between types of matches i.e. if match was played as FPP (First Person Perspective) or TPP (Third Person Perspective). We had also to drop some inconsistend variables, which were mentioned in the data set overview.

EDA

After data manipulation, we are left with around 14 continuous and binary variables, used further for creating models. All of them have good distributions, many displaying significant normality. Our area of interest were only single matches, so we have dropped at this point all observation that are not single-mode. Hence, variables like revives, teamKills are of no use for us anymore. Yet, we can leave them as they always show the same value (0).



Machine Learning Models

We decided to compare 4 machine learning classification methods – Logistic Regression, SVM, KNN and RandomForest. We started with running basic Logit Regression on all variables to get a glimpse into the data, remove some insignificant one, and check coefficients. As our prediction aim is to classify person, basing on their game results, as if he can reach with this stats top 25 – we have slightly unbalanced data set. 75% of our observations were 0 and 25% were 1, as our dependent variable is binary. Thus, we had to check two approaches. First, try to predict on unbalanced data with more complex Machine Learning method like KNN or Random Forest, hence Logit doesn't give any effects. Second, we performed undersampling of our bigger group, trained models on the undersampled data, and then made prediction on real, unbalanced data.

Findings

What we have found during this analysis was a bit of surprise. As we expected, classification on unbalanced data with Logit, gave us really bad results. When random guess to always bet that you won't be in top 25% of players is 75%, Logit gave us results of approximately 76% so only 1% better. Yet, KNN, SVM and RandomForest gave us better result, every of them gave us accuracy of approximately 80-81% (unfortunately when we consider unbalanced data, accuracy is not a good measure), thus we also investigated measures like Recall and Precision, to see exactly which method classifies our data best. It appeared that it's RandomForest.

Next Step of our analysis, was training our data on Balanced Sample. We did the undersampling, performed training, and then… it appeared that Logit Regression IMPROVED! From approx. 61% of accuracy when prediction, and train was made on unbalanced data to 80% of accuracy when we trained it on balanced data, and validated on real one. Yet, other models stayed the same, and RandomForest even worsen (unexpectedly).