ESSEC
BUSINESS SCHOOL

CentraleSupélec

Foundations of Machine Learning
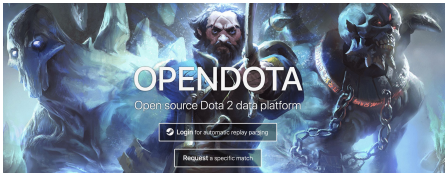## "Predicting the Outcome of a Dota 2 Game"

Arnaud Cluzel

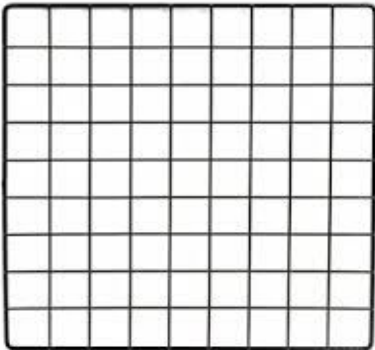**Victor Sanchez Chavez**

Julien Crabie

Valeriia Chirkova

Data Collection

Website Open Dota

OPENDOTA
Open source Dota 2 data platform

Dataset Build

API

Dota 2 API Loader

29 185
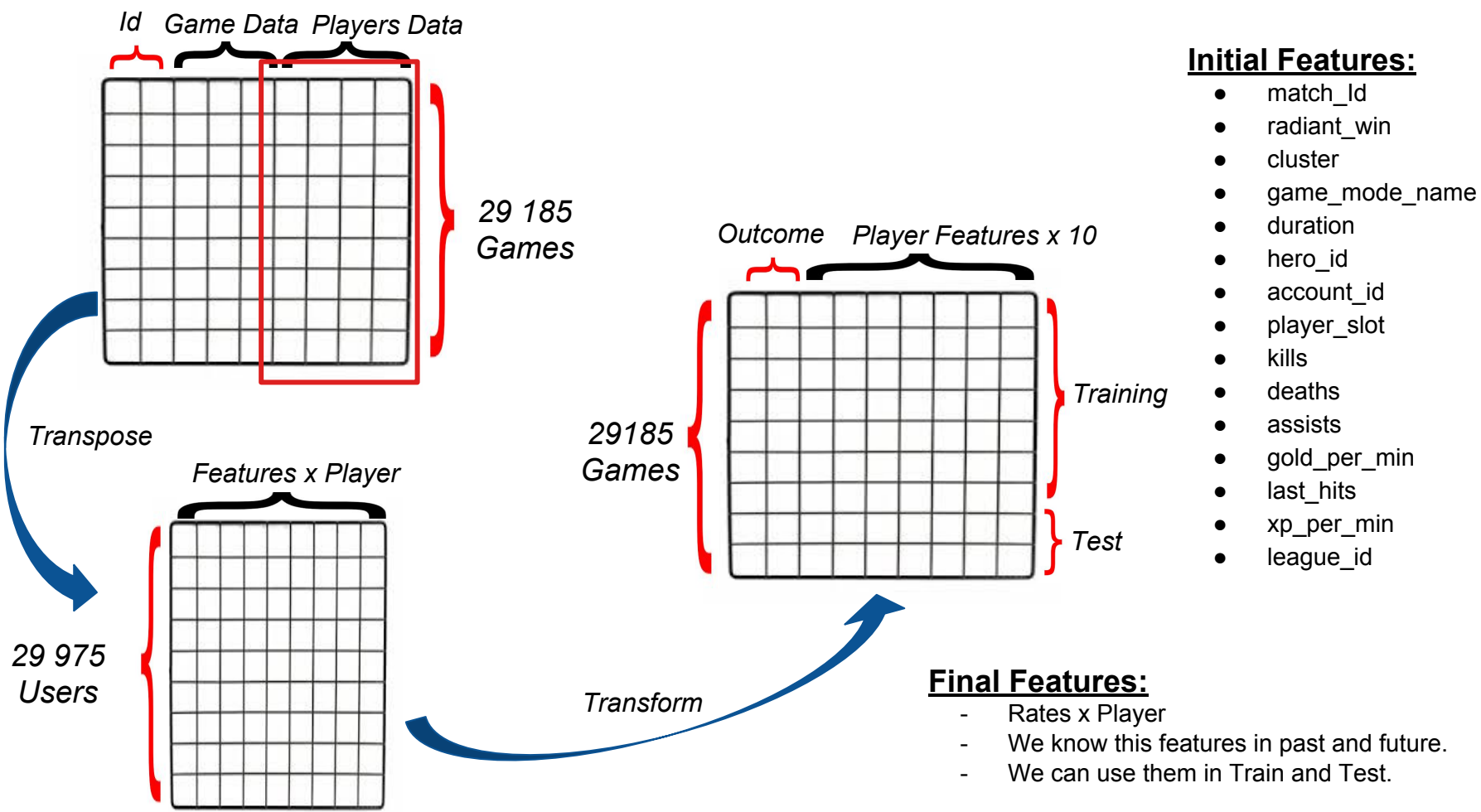Games

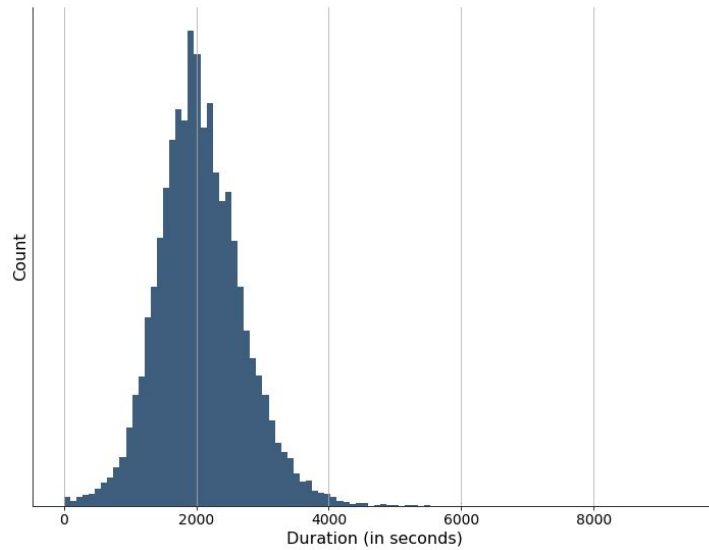*Manual List of **437** Professional players and their teams*

Constraints:
- Some users don't share their information (private profile).
- Some players in games are not sharing their account_Id. We get an "unknown" account_id (4294967295).
- Problems with some feature (game_mode_name).

Strategic Approach for our Prediction

Initial Features:
- match_Id
- radiant_win
- cluster
- game_mode_name
- duration
- hero_id
- account_id
- player_slot
- kills
- deaths
- assists
- gold_per_min
- last_hits
- xp_per_min
- league_id

Final Features:
- Rates x Player
- We know this features in past and future.
- We can use them in Train and Test.

Data Preparation and Data Visualization

Games duration

Dealing with **inconsistent** and **non-representative** values
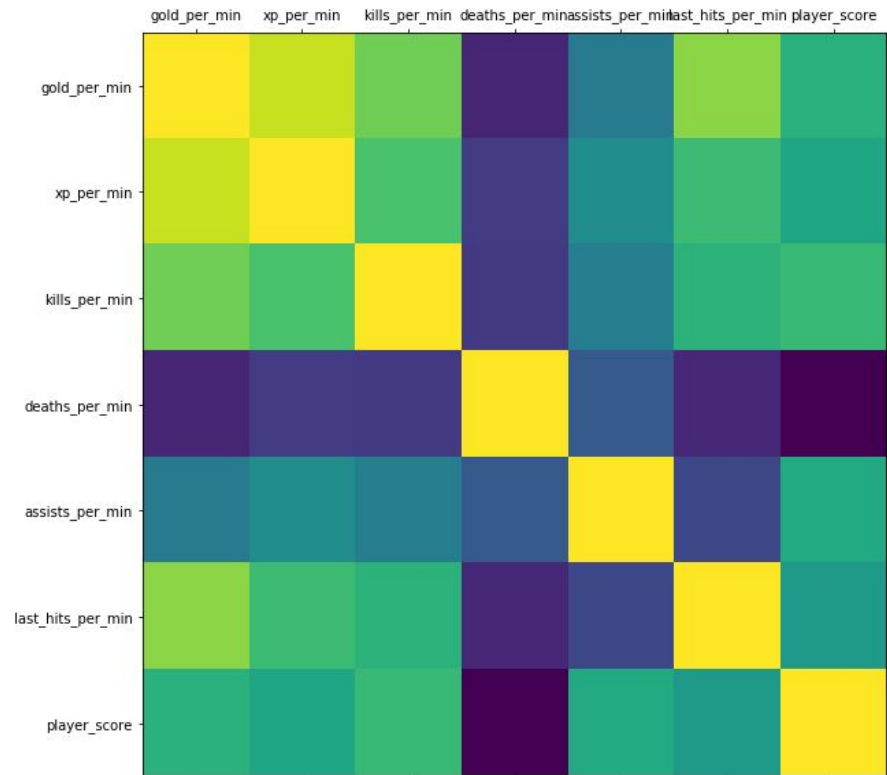
Dealing with **unknown players**
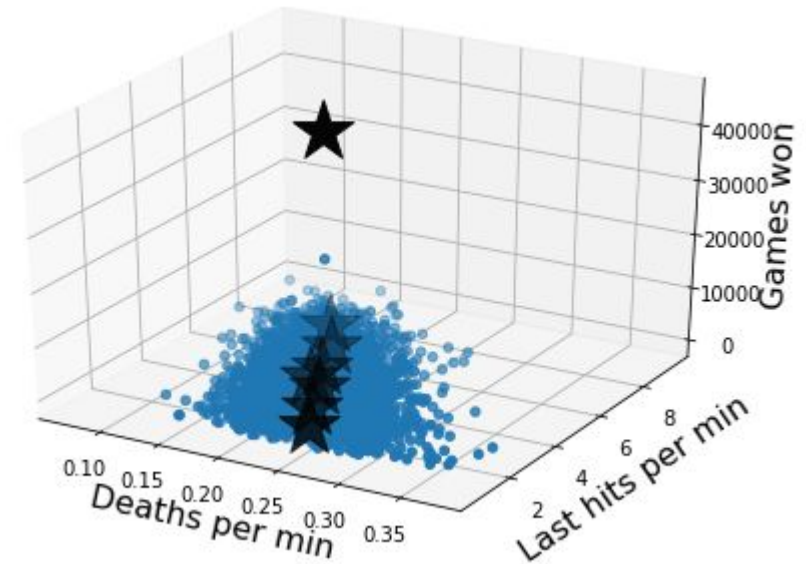
3D plot with unknown player Id

Feature engineering

Creation of new variables:

- Kills per min

- Deaths per min

- Last hits per min

- Assists per min

- Xp per min

- Gold per min
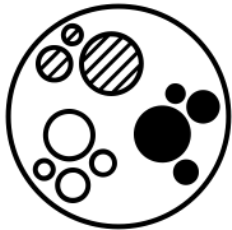
- Player score

Player metrics correlation plot

Model Training

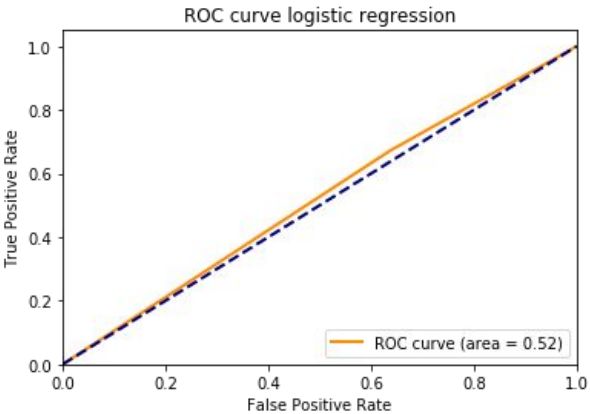Logistic regression

Support vector machine

Gaussian | Linear

SVD

Clustering

Feature selection

Evaluation

Accuracy scores

	Logistic regression	SVM (kernel = gaussian)		SVM (kernel = linear)
Train	71,9%	99,9%	59,8 % with CV	71,6%
Test	53,0%	58,5%		53,0%

ROC curve logistic regression

Logistic regression

	-1	+1
-1	409	718
+1	436	893

SVM

	-1	+1
-1	108	1019
+1	0	1329

ROC curve SVM Gaussian kernel

Conclusions
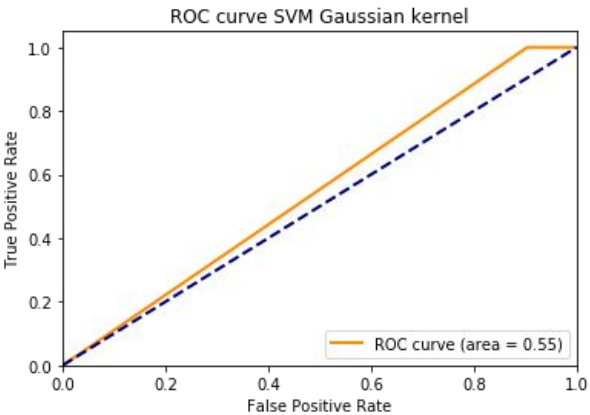
● Low Accuracy - High Variance.

● More Data required on players.

● Improvement for future work:
  ○ Double the dataset by switching teams' side.
  ○ Use more extensive data on players for better scoring.
  ○ Use more extensive data for a better player ordering.
  ○ Use more extensive data for the creation of teams features.