Topic: Data Analytics

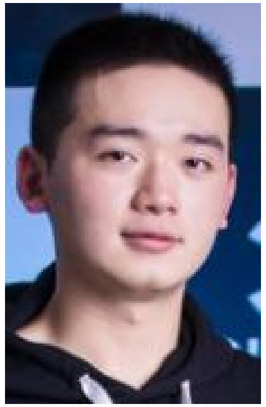
Research question:
In online games, predict whether a player is active or not in the next week based on the player’s previous behavior.

Aim 1: improve the model's accuracy.

Aim 2: analyse players' behaviors and give advice for keeping them.

Tutor: Tommy Liu

Group member:



Qingzheng Xu
u6174243

Xuguang Song
u6250082

This question is important because:

1. It can help game designers to make the game better, and they can know how to adjust strategy for different players. **Business value.**

2. Players’ in-game behaviours are also meaningful in helping us analyse real-life behaviours. For example, in 2007, some researchers analyse real-world epidemics based on a virtual epidemic in World of Warcraft. **Research value.**

Raw data resource (timestamp):

Original data comes from World of Warcraft, officially recorded at one of the server between 2008-2010.

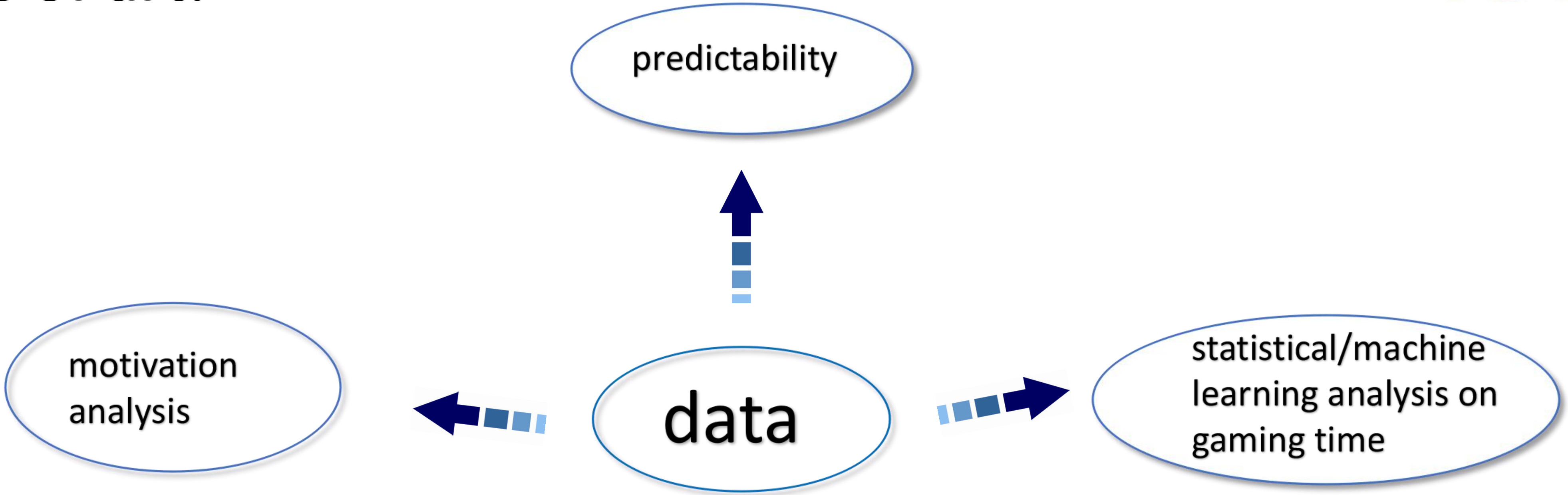
We use the open source version(License CC0: Public Domain) on Kaggle shared by Myles O'Neill (2019).

What the raw data set looks like:

	char	level	race	charclass	zone	guild	timestamp
0	59425	1	Orc	Rogue	Orgrimmar	165	01/01/08 00:02:04
1	65494	9	Orc	Hunter	Durotar	-1	01/01/08 00:02:04
2	65325	14	Orc	Warrior	Ghostlands	-1	01/01/08 00:02:04
3	65490	18	Orc	Hunter	Ghostlands	-1	01/01/08 00:02:04
500	24815	70	Orc	Warlock	Shattrath City	104	01/01/08 00:12:28
501	49719	70	Orc	Warlock	Netherstorm	79	01/01/08 00:12:28
502	27427	23	Tauren	Shaman	Undercity	160	01/01/08 00:12:33
503	60859	44	Tauren	Shaman	Warsong Gulch	228	01/01/08 00:12:33
1000	87	70	Tauren	Shaman	Nagrand	19	01/01/08 00:23:23
1001	30078	70	Tauren	Shaman	The Steamvault	5	01/01/08 00:23:23
1002	39365	70	Tauren	Shaman	Netherstorm	35	01/01/08 00:23:23
1003	60179	70	Tauren	Shaman	Arathi Basin	53	01/01/08 00:23:23

Recording all current online players' information every **10** minutes.

State of art:



Our choice:

define player groups and machine learning
+
player behavior analysis

Build data frame:

	char	race	charclass	meanlevel	maxlevel	minlevel	levelup	guild	zone	soloplayer	w1_active	12_change	w2_active	23_change	w3_active	34_change	w4_active	leave
0	7	1.0	2.0	57.30	60	54	6	282	117	0	0	0	0	172	172	-113	59	0
1	9	1.0	2.0	70.00	70	70	0	79	80	0	216	-19	197	-71	126	51	177	0
2	19	1.0	1.0	69.93	70	69	1	-1	60	0	6	160	166	-19	147	36	183	0
3	21	1.0	2.0	70.00	70	70	0	205	18	0	75	-34	41	-12	29	6	35	0
4	22	1.0	3.0	62.00	62	62	0	5	121	0	0	0	0	15	15	-15	0	1

Attributes:

char, race, charclass, meanlevel, maxlevel, minlevel, levelup

guild, zone, soloplayer

wi_active, ij_change: such as w1_active, 12_change

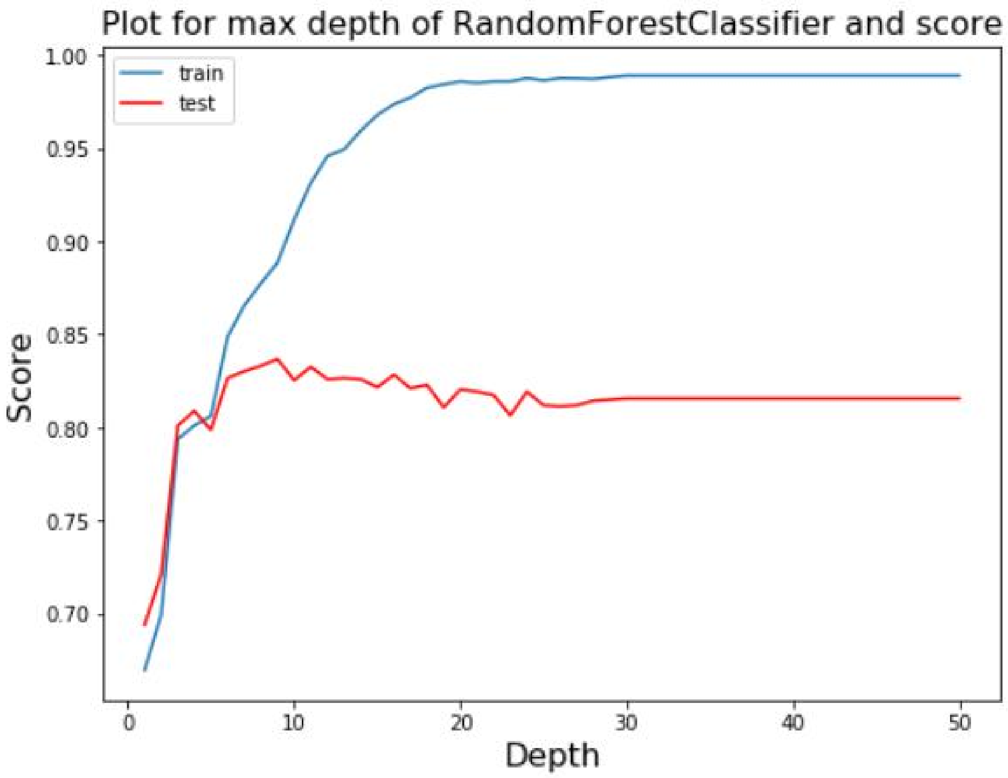
Target attribute:

Leave (0 or 1)

Train classification model and Overall accuracy:

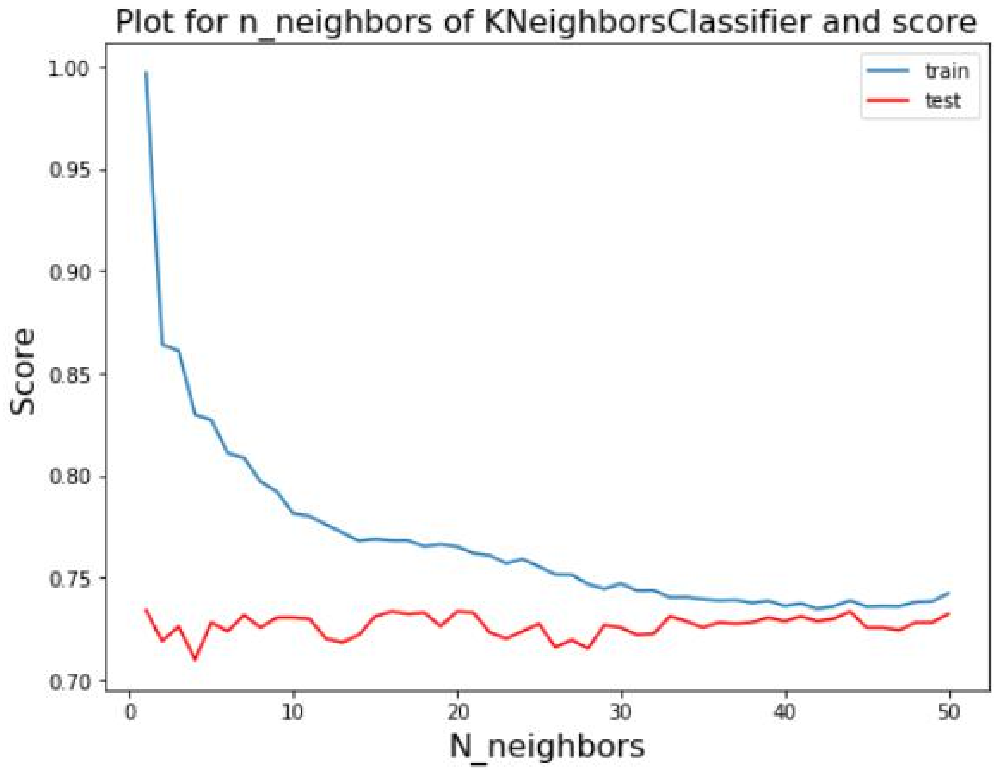
Accuracy: model.score(x_test, y_test)
The ability of the model to correctly predict the target attribute – leave.

83.68% for the best Random Forest model



The best max_depth is 9 and score is 0.8367593712212817 .

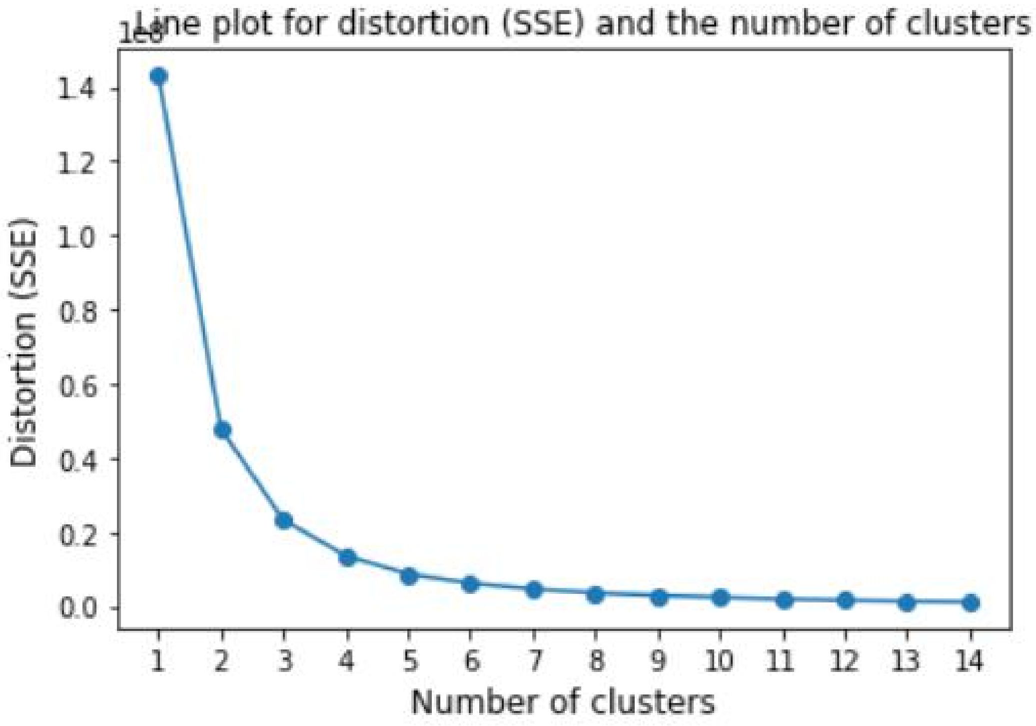
73.40% for the best K-Nearest Neighbors model

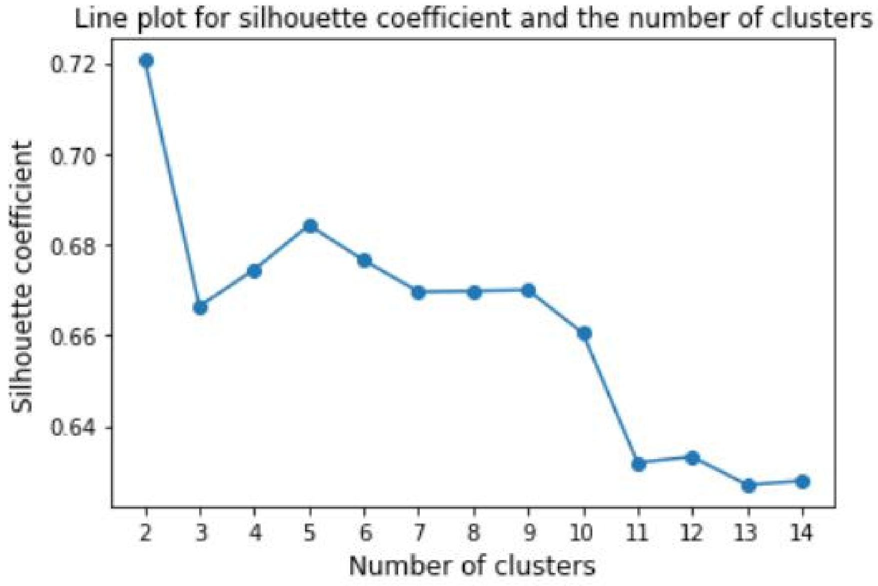


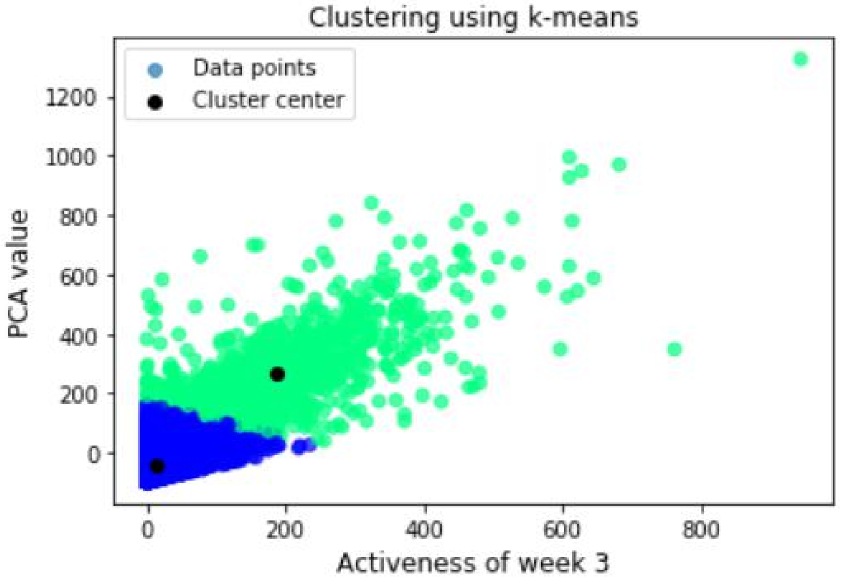
The best n is 1 and score is 0.7339782345828295 .

Approach 1: Clustering

Method for choosing the number of clusters:
Elbow method and silhouette coefficient







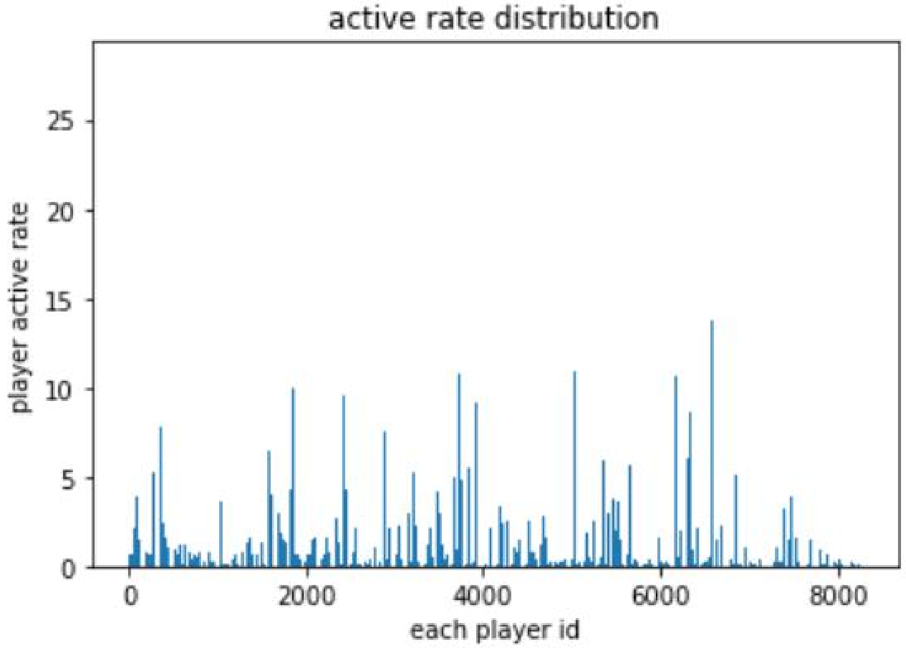
Approach 2: Manually divide

Definition: active rate =

$$\frac{\sum_{i=1}^{3} week_i\_active\_time + week1\_to\_2\_change\_time + week2\_to\_3\_change\_time + max\_level + level\_up}{100}$$

Distribution:

```
max active rate:  27.96
min active rate:  0.01
mean active rate:  1.443015239477533
```



Division:

Casual player: active rate < 0.1

Average player: 0.1 < active rate < 1.6

Loyal Player: active rate > 1.6

Outcomes of our project:

Australian National University

General model accuracy (using the whole dataset):

83.68% for the best Random Forest model. 73.40% for the best K-Nearest Neighbors model.

K-means clustering

outcome and overall accuracy of separate data sets:

	group1	group2	overall
RF classifier	95.89% (+22.49%)	81.74% (-1.93%)	83.61% (-0.07%)
KNN classifier	95.9% (+22.5%)	71.7% (-1.69%)	74.9% (+1.5%)

Manually dividing based on active rate

outcome and overall accuracy of separate data sets:

	casual	average	loyal	overall
RF classifier	90.25% (+6.58%)	77.90% (-5.8%)	95.71% (+12.03%)	84.53% (+0.85%)
KNN classifier	78.3% (+4.9%)	69.18% (-4.22%)	95.71% (+22.31%)	77.28% (+3.88%)

Conclusion:

1. Dividing the whole data into small groups containing different players can improve accuracy.

2. Different kinds of players have quite different behaviors, so the company should treat them differently and try to turn casual or average players into loyal players.

More analytics and Recommendations:

23.93% **Loyal players** : to find challenges

More instances; high rewards; guild experience.

19.21% **Casual players** : to have a try at the game

advertisements; game tutorials; beginner protection.

56.86% **Average players** : to have fun and to pass time

more game contents; ask for feedback.

Future work:

1. Improve the prediction accuracy on average players

* combine the casual and loyal player's data to predict
* further dividing the players in average players

2. Treat the attribute 'guild' in a better way

* use a historic guild list for recording guild changes

such as [(-1, 5), (225, 10), (-1, 2), (126, 50)]

Tips: '-1' means no guild; each timestamp is 10 minutes.

Reference:

Eric & Nina (2007) The untapped potential of virtual game worlds to shed light on real world epidemics Retrieved from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(07)70212-8/fulltext?code=lancet-site&version=printerFriendly