Predicting Player Spending in Brutal Age

Yunhan Liu(yl3287); Jiawei Gao(jg2352)

Introduction

Introduction

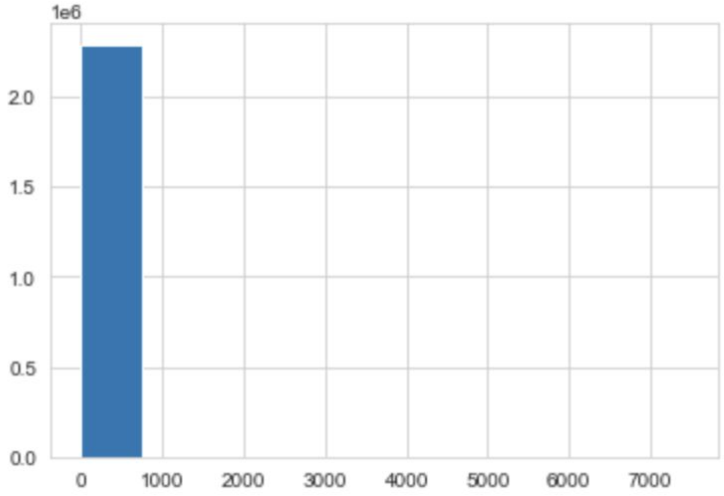
- Dataset resource: the second Intelligent China Cup (ICC) data competition
- Goal: estimate the total payment amount for players in the mobile game "Brutal Age" within the first 45 days of gameplay.
- Shape of Dataset: 2288007 rows, 109 columns, last column is our target variabel, payment_45
- Missing Value: No missing value
- Type of feature: 107 numeric features and 1 categorical feature(register_date)

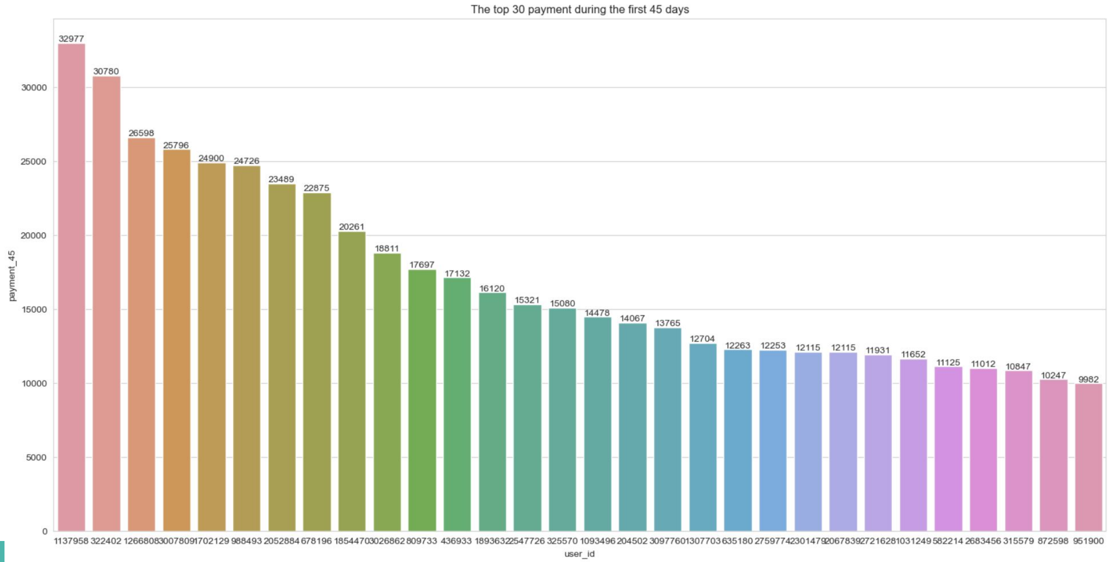
Data Exploration

Data Exploration

Histogram of payment in 45 days V.S Top 30 payment in 45 days

```
Rate of payment after trial: 2.010%
Number of people who had paid: 45988
Total payment of the player: 4102730.110
ARPU: 1.793
ARPPU: 89.213
Proportion of the amount spent by the top 500 paying users: 51.618%
Proportion of the amount spent by the top 1000 paying users: 64.878%
Proportion of the amount spent by the top 5000 paying users: 89.375%
```





Data Exploration

The data is unbalanced distributed based on different month, then we guess the month feature may matter

	Number of User	Paid User	Rate of payment	Total payment	Max_Payment
Month 1	390420	9220	2.362%	735105.930	24726.51
Month 2	1632463	30359	1.860%	2855919.480	32977.81
Month 3	265124	6409	2.417%	511704.700	18811.66

Feature Engineering

Feature Engineering

- ➢ Convert the "register_time" column to datetime format
- ➢ Add a new column called "month" with categorical data (values 1, 2, and 3 representing the month of registration)
- ➢ Delete the original column "register_time"

Modeling Work

Linear regression model

Mean squared error: 3913.130483320636

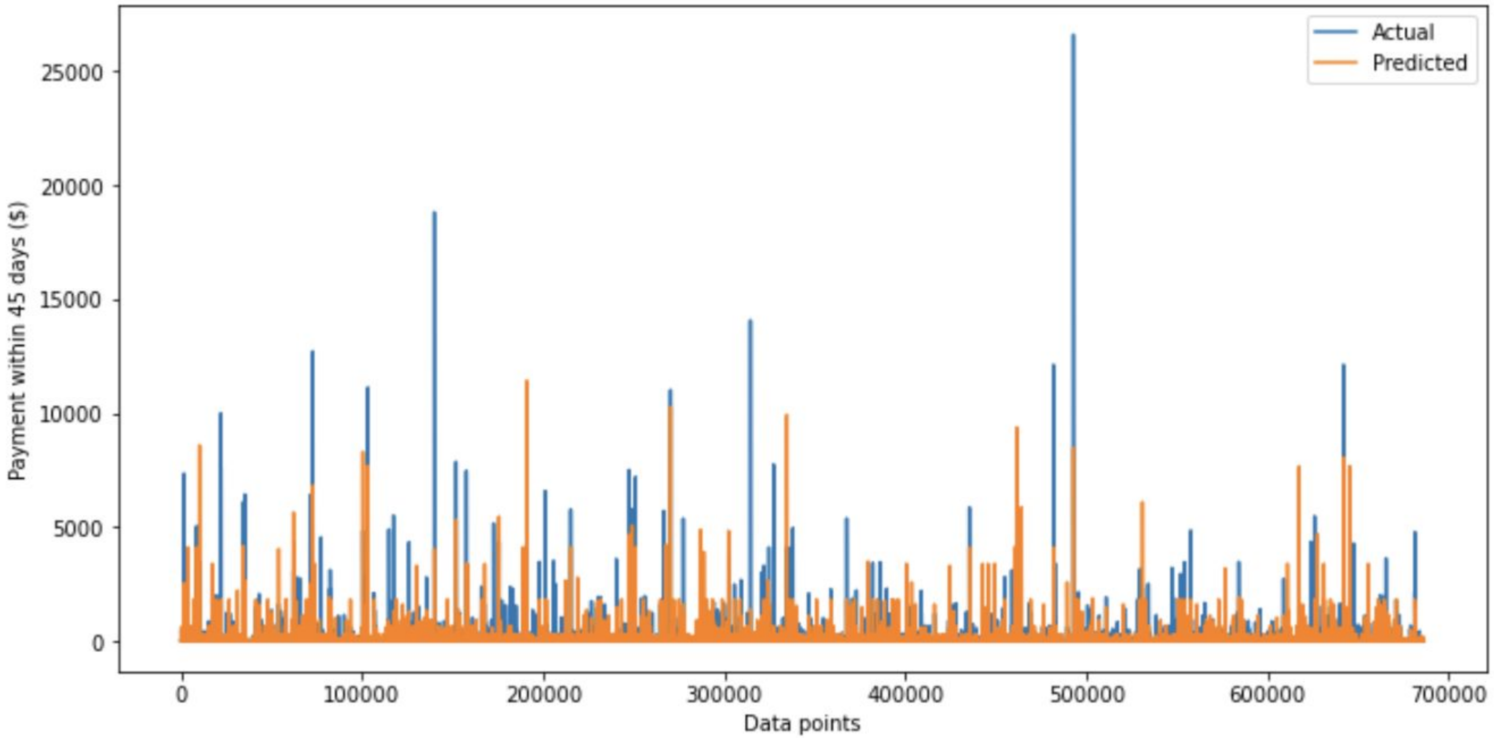
Random Forest

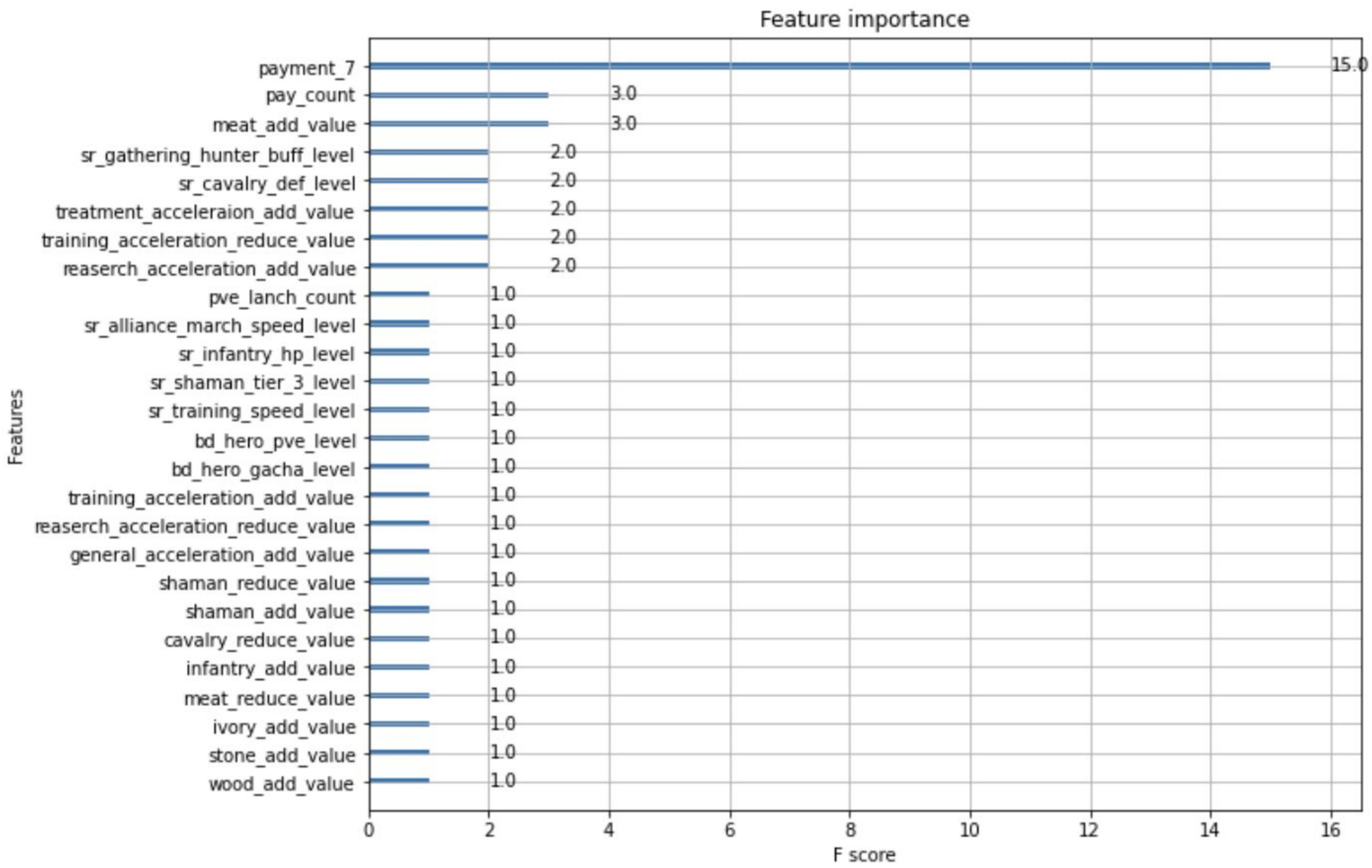
Model	Split	n_estimator	max_depth	MSE	R^2 on train	R^2 on test
m1	80%, 20%	100(default)	6(default)	4076.5398	0.8966	0.3149
m2	70%, 30%	10	6(default)	3873.5419	0.8981	0.3733
m3	70%, 30%	10	5	3424.0279	0.7576	0.4461
m4	70%, 30%	10	4	3189.5639	0.7163	0.4840
m5	70%, 30%	10	3	**3176.2768**	0.6514	0.4861
m6	70%, 30%	8	3	3188.3149	0.6475	0.4842

XGBoost

Model	Split	n_estimator	max_depth	MSE	R^2 on train	R^2 on test
m1	80%, 20%	10	6(default)	3975.6570	0.8813	0.3319
m2	70%, 30%	10	6(default)	4045.7498	0.8877	0.3454
m3	70%, 30%	10	3	3024.4583	0.7528	0.5107
m4	70%, 30%	10	2	3098.5843	0.6452	0.4987
m5	70%, 30%	8	3	3005.7318	0.7306	0.5137
m6	70%, 30%	7	3	2965.5106	0.7154	0.5202

XGBoost Visualization




Feature importance
payment_7
pay_count
meat_add_value
sr_gathering_hunter_buff_level
sr_cavalry_def_level
treatment_acceleraion_add_value
training_acceleration_reduce_value
reaserch_acceleration_add_value
pve_lanch_count
sr_alliance_march_speed_level
sr_infantry_hp_level
sr_shaman_tier_3_level
sr_training_speed_level
bd_hero_pve_level
bd_hero_gacha_level
training_acceleration_add_value
reaserch_acceleration_reduce_value
general_acceleration_add_value
shaman_reduce_value
shaman_add_value
cavalry_reduce_value
infantry_add_value
meat_reduce_value
ivory_add_value
stone_add_value
wood_add_value
15.0
3.0
3.0
2.0
2.0
2.0
2.0
2.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
0
2
4
6
8
10
12
14
16
F score
Features

Conclusion

In conclusion, based on the evaluation metrics, the XGBoost model performed the best among the models we built. This model can be used to predict the likelihood of a customer paying within 45 days, and the feature importances can provide insights to the business about which factors are most important for predicting customer payment behavior.

Improvement: build models with most important features