

MAINGAMES:

DATA SCIENCE CHALLENGE

Created by:

Muhamad Rifki Taufik




PREFACE AND OUTLINE

The written report is developed for answering the Data Science Challenge by Main Games. The given dataset consists of more than 600 streamers (Indonesia, Vietnam, Philippines). Each row represents data of one unique streamer. All end-to-end analysis process (import, data wrangling, EDA, visualization, analysis) utilized R programming with several libraries. The report outlines are:

1. Introduction
2. EDA-Descriptive Analysis
 - a. Table EDA
 - b. MBTI Grouped vs Gender
 - c. Games vs Gender
3. Correlation
 - a. Ranked Cross-Correlations
 - b. The dependent variable against all
4. Machine Learning Approach
5. Conclusion

Disclaimer: The writer does not have the detail description of dataset and the interpretation is based on model result.



1. Introduction

The dataset consisted of 631 observations and 127 variables where the task mentioned the PaidStarPerWatchedHour. This dataset is quite challenges because we have hundreds of variables with only 631 observations. The writer have tried several machine learning algorithms, however the dataset seems unsuitable for few models due to several issues. The issues might impact to the analysis, such as Neural Networks with nnet package did not able to create a hundreds-complex network and Linear Regression literation takes forever to find the best fitted model. Moreover, the max value for each variable should be given to get better insight and detect anomalies. The data scientist should be given by knowledge regarding dataset. In this problem, we want to learn whether **certain features** of streamers will make them more or less likely to receive stars from their audience.

2. EDA-Descriptive Analysis

a. Table

Since we have hundreds of variable this table is provided to get some idea how the data distributed. Tiny proportion of NA were found in Gender, Game, PaidPerWatchHour, and Follower. The DS simply omit the NA since the NA around 1% or less. Majority the players come from Indonesia and Vietnam and their most favorite games are MLBB and PUBG. Most of the players are male.

Variable	NotNA	Mean	Median	PropNA
Country	631			
... ID	232	36.8%		
... PH	159	25.2%		
... VN	240	38%		
Gender	628			3 (0.47%)
... Female	175	27.9%		
... Male	453	72.1%		
Game	624			7 (1.01%)
... 8 Ball Pool	2	0.3%		
... Age of Empires	26	4.2%		
... Agge of Empires	2	0.3%		
... Apex Legends	3	0.5%		
... Arena of Valor	42	6.7%		
... Assassin's Creed Odyssey	1	0.2%		

Variable	NotNA	Mean	Median	PropNA
... Audition	1	0.2%		
... Audition Online	1	0.2%		
... Auto Chess	4	0.6%		
... Blade & Soul	1	0.2%		
... CABAL ONLINE	2	0.3%		
... Call of Duty: Mobile	2	0.3%		
... Call of Duty: Mobile VN	3	0.5%		
... Call of Duty: Warzone	2	0.3%		
... Coin Master	7	1.1%		
... Counter-Strike: Global Offensive	8	1.3%		
... Crazy Kart	1	0.2%		
... Crossfire	13	2.1%		
... Days Gone	2	0.3%		
... Dead by Daylight	4	0.6%		
... Dota 2	14	2.2%		
... EA Sports UFC 3	1	0.2%		
... eFootball PES 2020	6	1%		
... Euro Truck Simulator 2	8	1.3%		
... FIFA Online 4	4	0.6%		
... Five Nights at Freddy's 2	1	0.2%		
... Free Fire - Battlegrounds	49	7.9%		
... Garena LiÃn QuÃn Mobile	12	1.9%		
... God of War	1	0.2%		
... Grand Theft Auto V	18	2.9%		
... Green Hell	1	0.2%		
... Identity V	3	0.5%		
... J-League Jikkyou Winning Eleven	1	0.2%		
... League of Legends	40	6.4%		
... League of Legends: Wild Rift	1	0.2%		
... LiÃn QuÃn Mobile	2	0.3%		
... Little Big Snake	1	0.2%		
... Minecraft	1	0.2%		
... MLBB	151	24.2%		
... Moon of Madness	1	0.2%		
... MotoGP	2	0.3%		
... MU Online	3	0.5%		
... Naruto Shippuden: Ultimate Ninja Storm 4	2	0.3%		
... NBA 2K20	3	0.5%		
... Ngá»• c Rá»ng Online	1	0.2%		
... No MLBB Video	1	0.2%		
... Persona 5 Royal	1	0.2%		
... Point Blank Indonesia	1	0.2%		

Variable	NotNA	Mean	Median	PropNA
... PUBG	118	18.9%		
... Ragnarok M: Eternal Love	1	0.2%		
... Roblox	1	0.2%		
... RULES OF SURVIVAL	4	0.6%		
... Sea of Thieves	3	0.5%		
... Star Wars Jedi: Fallen Order	1	0.2%		
... Teamfight Tactics	4	0.6%		
... The Last of Us	1	0.2%		
... The Last of Us: Part II	1	0.2%		
... The Warriors	1	0.2%		
... Township Mobile	1	0.2%		
... Valorant	9	1.4%		
... World War Z	1	0.2%		
... Wormate.io	8	1.3%		
... Worms Zone	12	1.9%		
... Yakuza: Kiwami 2	1	0.2%		
Total.Follower	627	220795.061	75565	0.006
Broadcast.Hours	631	127.87	113	0
PaidStarPerWatchedHour	629	0.012	0.003	0.003
Character_Facet_Cont_Rigidity	631	0.602	0.593	0
Character_Facet_Cont_AchievementStriving	631	0.511	0.524	0
Personal_Values_Facet_Cont_Hedonism	631	0.505	0.539	0
Character_Facet_Cont_Dutifulness	631	0.496	0.506	0
Character_Cont_Conscientiousness	631	0.475	0.489	0
Character_Facet_Cont_Sympathy	631	0.398	0.389	0
Personal_Values_Facet_Cont_UniversalismTolerance	631	0.558	0.553	0
Temperament_Choleric	631	0.192	0.25	0
Self_Esteem_Cont_SEDiscrepancyResponsibility	631	0.502	0.537	0
Character_Facet_Cont_ArtisticInterests	631	0.538	0.538	0
Temperament_Sanguine	631	0.333	0.25	0
Character_Facet_Cont_Friendliness	631	0.521	0.539	0
Temperament_Diligent	631	0.327	0.375	0
Self_Esteem_Cont_SELevel	631	0.555	0.559	0
Character_Cont_Extraversion	631	0.482	0.478	0
Role_Director	631	0.523	0.531	0
Character_Facet_Cont_SelfConsciousness	631	0.481	0.493	0
Temperament_Energetic	631	0.326	0.25	0
Character_Facet_Cont_Adventurousness	631	0.603	0.619	0
Role_Craftsman	631	0.476	0.504	0
Self_Esteem_Cont_SELevelResponsibility	631	0.662	0.671	0
Self_Esteem_Cont_SEAspirationResponsibility	631	0.713	0.726	0
Role_Toastmaster	631	0.51	0.544	0

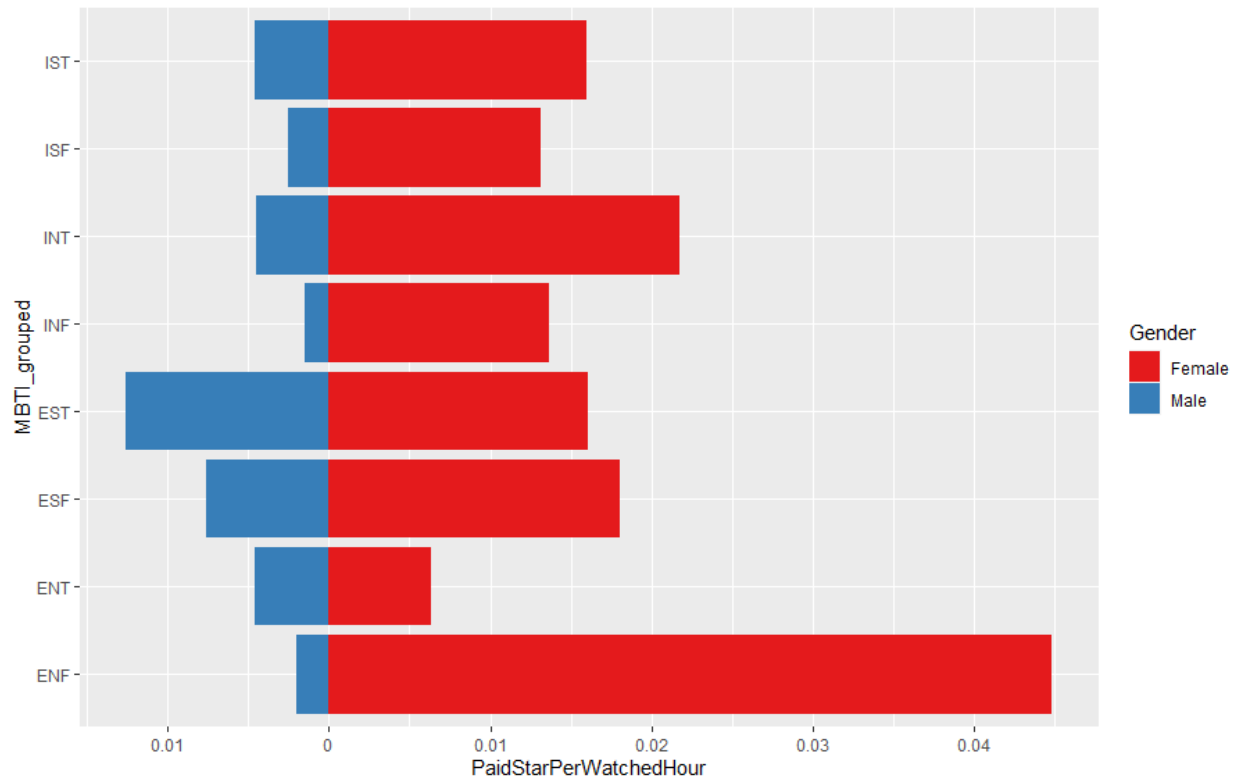
Variable	NotNA	Mean	Median	PropNA
Personal_Values_Facet_Cont_PowerDominance	631	0.344	0.343	0
Role_Marshal	631	0.55	0.542	0
Self_Esteem_Cont_SELevelAchievements	631	0.428	0.457	0
Self_Esteem_Cont_SEDiscrepancySociability	631	0.504	0.504	0
Personal_Values_Cont_SelfEnhancement	631	0.277	0.291	0
Self_Esteem_Cont_SEAspiration	631	0.866	0.876	0
Role_Supplier	631	0.355	0.353	0
Character_Facet_Cont_Activity	631	0.591	0.575	0
Temperament_Stable	631	0.337	0.375	0
Personal_Values_Facet_Cont_PowerResources	631	0.277	0.313	0
Self_Esteem_Cont_SEDiscrepancyAchievements	631	0.538	0.544	0
Temperament_Melancholic	631	0.193	0.25	0
Character_Cont_Openness	631	0.583	0.583	0
Character_Facet_Cont_Anxiety	631	0.579	0.573	0
Role_Administrator	631	0.511	0.513	0
Character_Facet_Cont_Emotionality	631	0.55	0.545	0
Role_Innovator	631	0.62	0.616	0
Role_Partygoer	631	0.447	0.446	0
Self_Esteem_Cont_SEAspirationSociability	631	0.763	0.762	0
Role_Guru	631	0.49	0.485	0
Personal_Values_Facet_Cont_SelfDirectionThought	631	0.28	0.27	0
Temperament_Phlegmatic	631	0.334	0.25	0
Personal_Values_Facet_Cont_Reputation	631	0.369	0.361	0
Role_Guardian	631	0.517	0.5	0
Role_Keeper	631	0.36	0.341	0
Character_Facet_Cont_Tension	631	0.568	0.565	0
Character_Facet_Cont_Altruism	631	0.344	0.354	0
Personal_Values_Facet_Cont_SecuritySocietal	631	0.435	0.448	0
Self_Esteem_Cont_SEAspirationAchievements	631	0.819	0.851	0
Self_Esteem_Cont_SEAspirationOriginality	631	0.763	0.765	0
Personal_Values_Facet_Cont_Humility	631	0.332	0.332	0
Character_Facet_Cont_Orderliness	631	0.43	0.415	0
MBTI_grouped	631			
... ENF	9	1.4%		
... ENT	279	44.2%		
... ESF	4	0.6%		
... EST	9	1.4%		
... INF	11	1.7%		
... INT	144	22.8%		
... ISF	5	0.8%		
... IST	170	26.9%		
Role_Manager	631	0.369	0.387	0

Variable	NotNA	Mean	Median	PropNA
Personal_Values_Facet_Cont_Stimulation	631	0.505	0.502	0
Character_Cont_Neuroticism	631	0.512	0.515	0
Self_Esteem_Cont_SELevelConfidence	631	0.514	0.54	0
Self_Esteem_Cont_SEAspirationConfidence	631	0.792	0.797	0
Role_WiseMan	631	0.5	0.514	0
Self_Esteem_Cont_SEDiscrepancyConfidence	631	0.529	0.526	0
Role_Assistant	631	0.332	0.338	0
Self_Esteem_Cont_SEDiscrepancy	631	0.456	0.443	0
Personal_Values_Facet_Cont_ConformityRules	631	0.509	0.519	0
Self_Esteem_Cont_SEAspirationIntelligence	631	0.795	0.792	0
Self_Esteem_Cont_SELevelSociability	631	0.725	0.766	0
Role_Healer	631	0.488	0.51	0
Role_Operator	631	0.527	0.552	0
Self_Esteem_Cont_SEDiscrepancyAppearance	631	0.497	0.482	0
Personal_Values_Facet_Cont_ConformityInterpersonal	631	0.55	0.543	0
Character_Facet_Cont_Modesty	631	0.604	0.62	0
Self_Esteem_Cont_SEDiscrepancyOriginality	631	0.562	0.56	0
Character_Facet_Cont_Imagination	631	0.518	0.514	0
Role_Coach	631	0.246	0.217	0
Role_RightsDefender	631	0.5	0.525	0
Role_Philanthropist	631	0.436	0.465	0
Personal_Values_Facet_Cont_SelfDirectionAction	631	0.526	0.538	0
Role_Promoter	631	0.555	0.58	0
Personal_Values_Facet_Cont_SecurityPersonal	631	0.47	0.462	0
Personal_Values_Facet_Cont_UniversalismConcern	631	0.501	0.486	0
Role_Advisor	631	0.442	0.453	0
Self_Esteem_Cont_SELevelOriginality	631	0.692	0.7	0
Role_Analyst	631	0.59	0.598	0
Personal_Values_Cont_SelfTranscendence	631	0.258	0.263	0
Personal_Values_Facet_Cont_Achievement	631	0.31	0.315	0
Character_Facet_Cont_Trust	631	0.47	0.491	0
Temperament_Unstable	631	0.196	0.125	0
Character_Cont_Agreeableness	631	0.32	0.318	0
Role_Inventor	631	0.625	0.676	0
Personal_Values_Facet_Cont_BenevolenceCaring	631	0.497	0.505	0
Personal_Values_Facet_Cont_BenevolenceDependability	631	0.3	0.308	0
Character_Facet_Cont_Vulnerability	631	0.68	0.678	0
Character_Facet_Cont_ExcitementSeeking	631	0.586	0.597	0
Role_Designer	631	0.57	0.621	0
Role_Deputy	631	0.318	0.326	0
Personal_Values_Cont_OpennessToChange	631	0.298	0.301	0
PaidStarPerWatchedHour.1	629	0.012	0.003	0.003

Variable	NotNA	Mean	Median	PropNA
Character_Facet_Cont_Morality	631	0.392	0.349	0
Character_Facet_Cont_Depression	631	0.636	0.662	0
Role_Strategist	631	0.628	0.636	0
Role_Inspector	631	0.532	0.512	0
Self_Esteem_Cont_SEAspirationAppearance	631	0.737	0.718	0
Character_Facet_Cont_SelfDiscipline	631	0.476	0.462	0
Self_Esteem_Cont_SEDiscrepancyIntelligence	631	0.425	0.472	0
Role_Companion	631	0.415	0.43	0
Character_Facet_Cont_Cooperation	631	0.385	0.369	0
Role_Commander	631	0.484	0.46	0
Role_Activist	631	0.419	0.439	0
Personal_Values_Facet_Cont_UniversalismNature	631	0.486	0.471	0
Role_Charismatic	631	0.483	0.475	0
Personal_Values_Facet_Cont_Tradition	631	0.354	0.342	0
Self_Esteem_Cont_SELevelAppearance	631	0.629	0.66	0
Character_Facet_Cont_Assertiveness	631	0.498	0.497	0
Character_Facet_Cont_Cautiousness	631	0.509	0.512	0
Role_Curator	631	0.265	0.247	0
Personal_Values_Cont_Conservation	631	0.299	0.301	0
Self_Esteem_Cont_SELevelIntelligence	631	0.77	0.776	0
Temperament_Centric	631	0.331	0.25	0

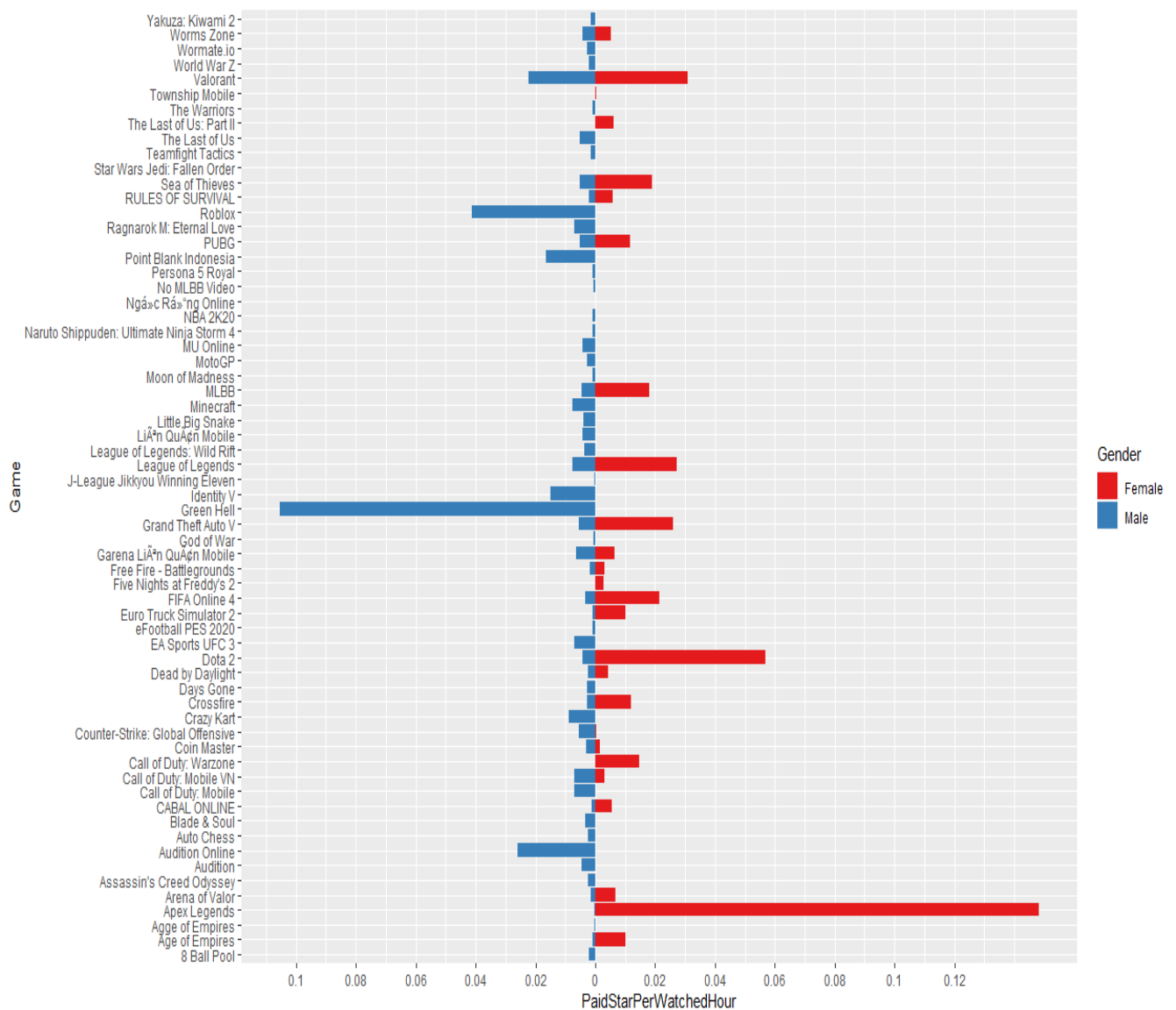
b. MBTI Group vs Gender

Obviously the female players dominated on PaidStarPerWatchHour, where the highest rank is in ENF group. The lowest PaidStarPerWatchHour Male players is in INF and ENF group. This is an interesting fact that in every group, the Female players mostly have higher PaidStarPerWatchHour.



c. MBTI Group vs Gender

Games always be the prior player's interest. According to the graph below, there are several insight that maybe we can identify whether games interested by male player only, more likely female player, or both. Valorant, Worm Zone, Garena becoming games played by both gender. However, female players have the upper hand to get PaidStarPerWatchHour for Valorant, Sea of Thieves, PUBG, MLBB, League of Legends, and Dota 2. Several games are played by male players only and some others female players only. Overall, the games played by both gender give female player more PaidStarPerWatchHour rather than male players.



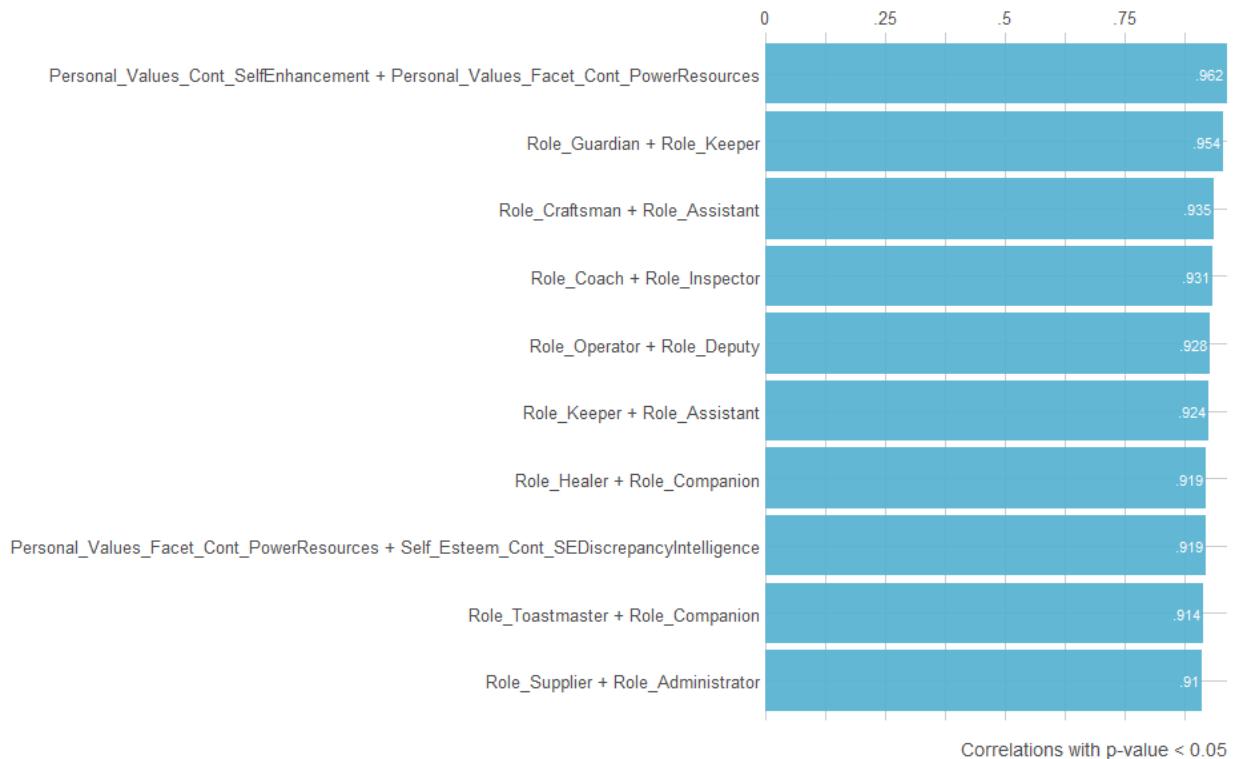
3. Correlation

a. Ranked Cross-Correlations

Cross correlations indicated that paired variable which have a significant association. Since we have hundreds of features, the graph below only shows the top ten of cross correlation with significant correlation (p value less than 0.05).

Ranked Cross-Correlations

10 most relevant



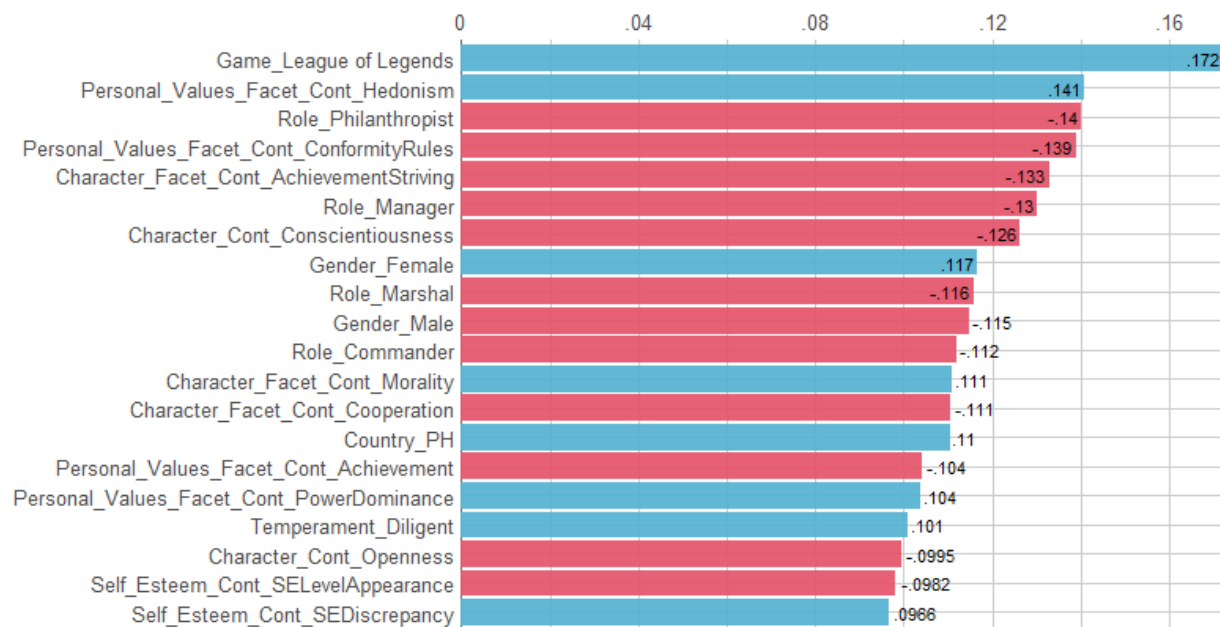
b. The dependent variable against all

Since our interest variable is PaidStarPerWatchHours, then we would like to identify the strongest paired correlations with it. There are positive and negative correlations where the positive correlation indicate if the paired variable increase then the PaidStarPerWatchHours also will increase. Oppositely, if the paired variable has negative correlation then if it increases means the PaidStarPerWatchHours will be decreasing. The strongest positive correlations against PaidStarPerWatchHours is Game League of Legends, follows Personal Value Facet Cont Hedonism. Moreover, the strongest negative correlations against PaidStarPerWatchHours is the

Role Philanthropist and Personal Values Facet Cont ConformityRules. These paired variables can be used to optimize the PaidStarPerWatchedHour with increase player with positive correlation character and reduce the player with negative correlation character.

Correlations of PaidStarPerWatchedHour

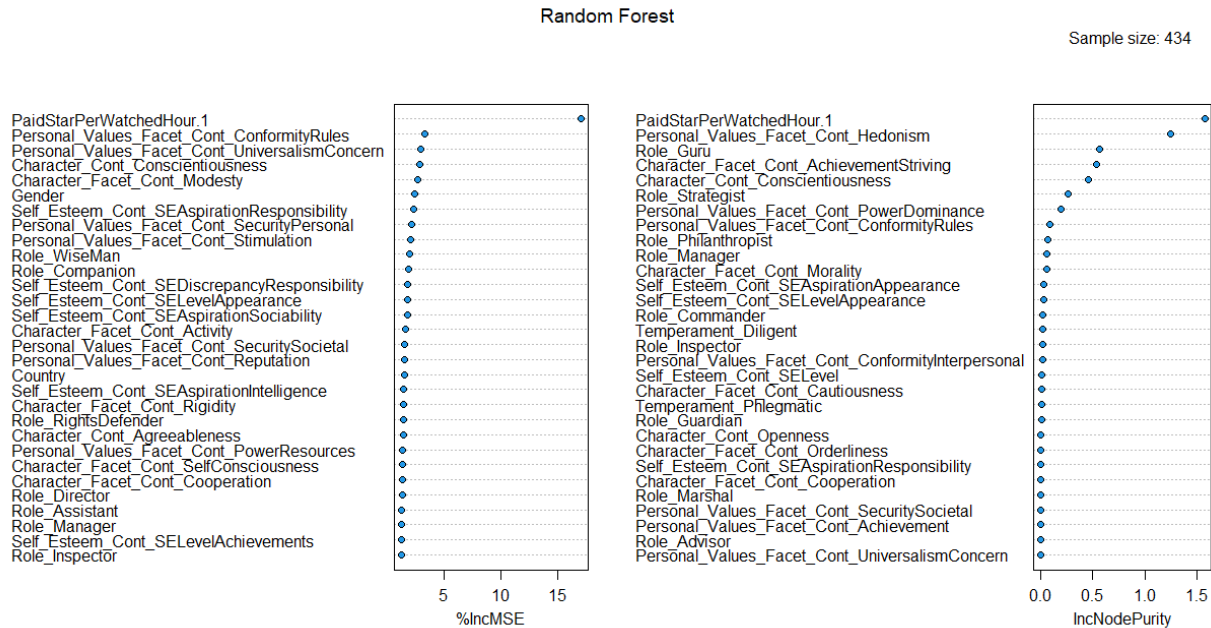
Top 20 out of 145 variables (original & dummy)



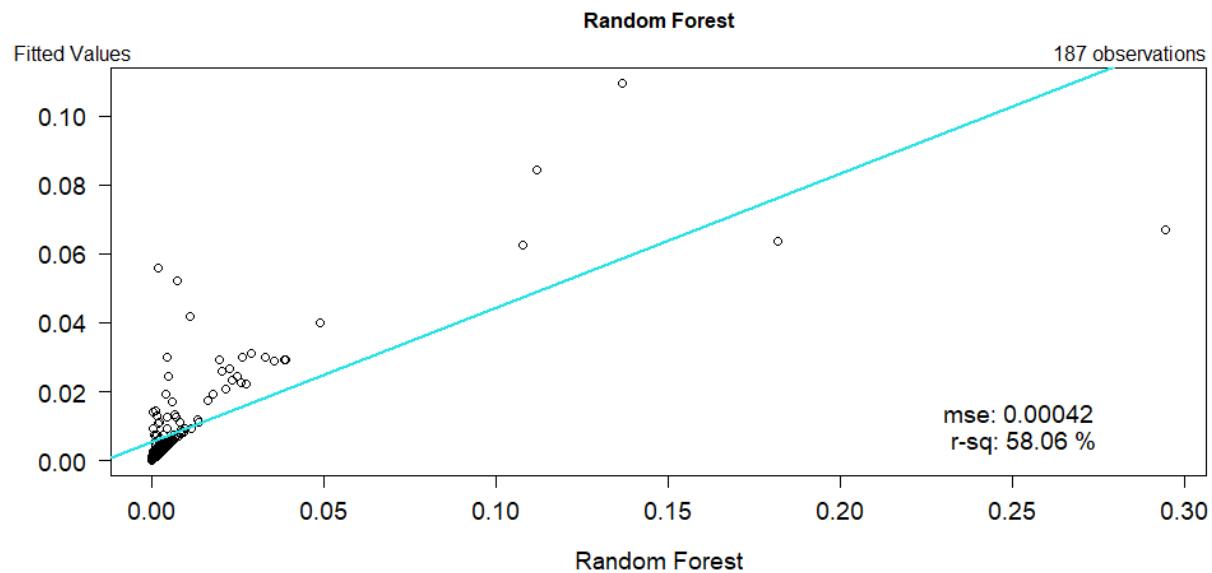
4. Machine Learning Approach

a. Random Forest

To identify the most important features among hundreds independent variables, we employs several machine learning algorithm, where the R square will evaluate the best model. The first algorithm is random forest (RF). The RF ranks the most important features regarding PaidStarsPerWatchedHour. The most important feature is PaidStarPerWatchedHour.1, the author does not know the differences between PaidStarPerWatchedHour.1 and PaidStarPerWatchedHour. It seems both features have a certain relationship. Moreover, important features to predict the PaidStarPerWatchedHour are Personal Value Facer Cont ConformityRules, UniversalismConcern, Character Cont Conscientiousness, character Facet Cont Modesty, Gender.



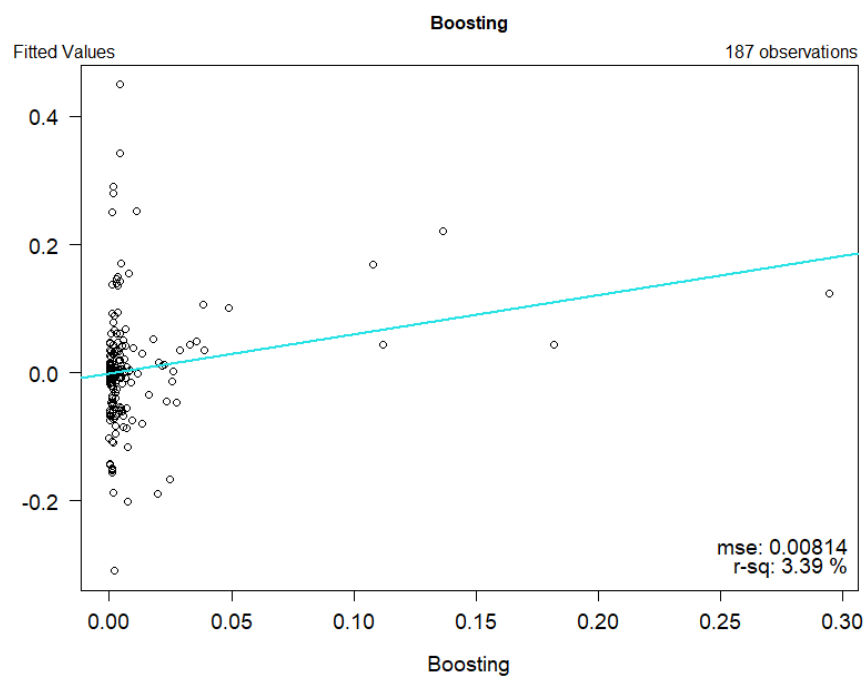
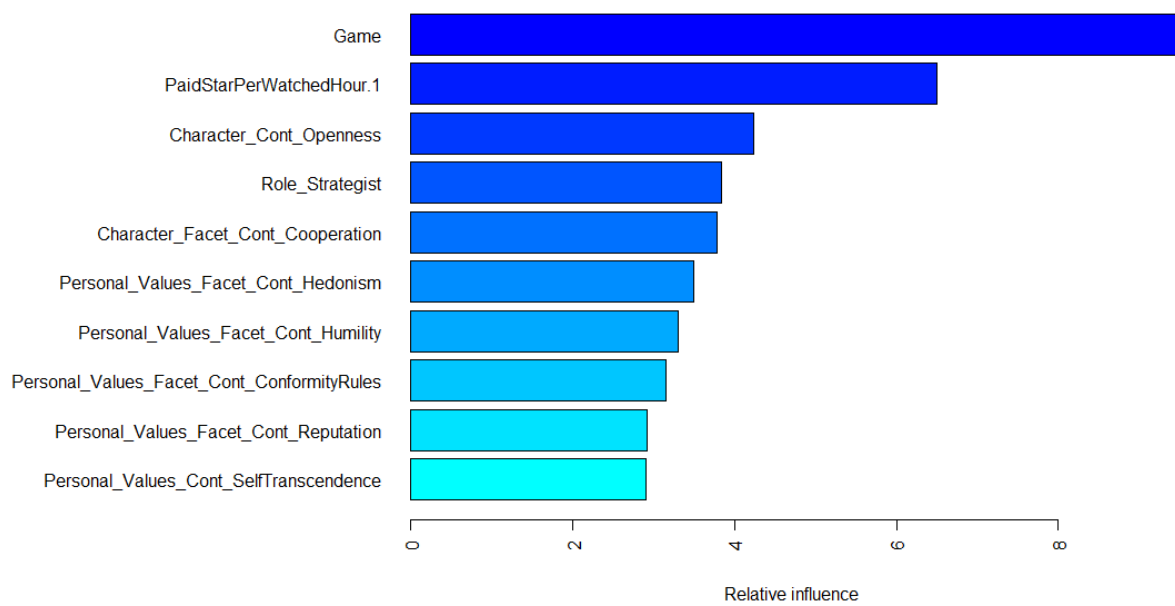
Based on these important feature, we tried to assess the model by predicting the testing set. The fitted value and observe values are depicted the following graph. According to the prediction, we evaluate the model using Mean Square Error (MSE) and R-Square. Algorithm with lowest MSE and the highest R square is the best model. Random Forest got MSE 0.00042 with R square 58%. The model with r square 58% indicates that the features can predict and explain the dependent PaidStarPerWatchedHour for 58%. This model is quite high and can be improved for getting better predictions.



b. Boosting

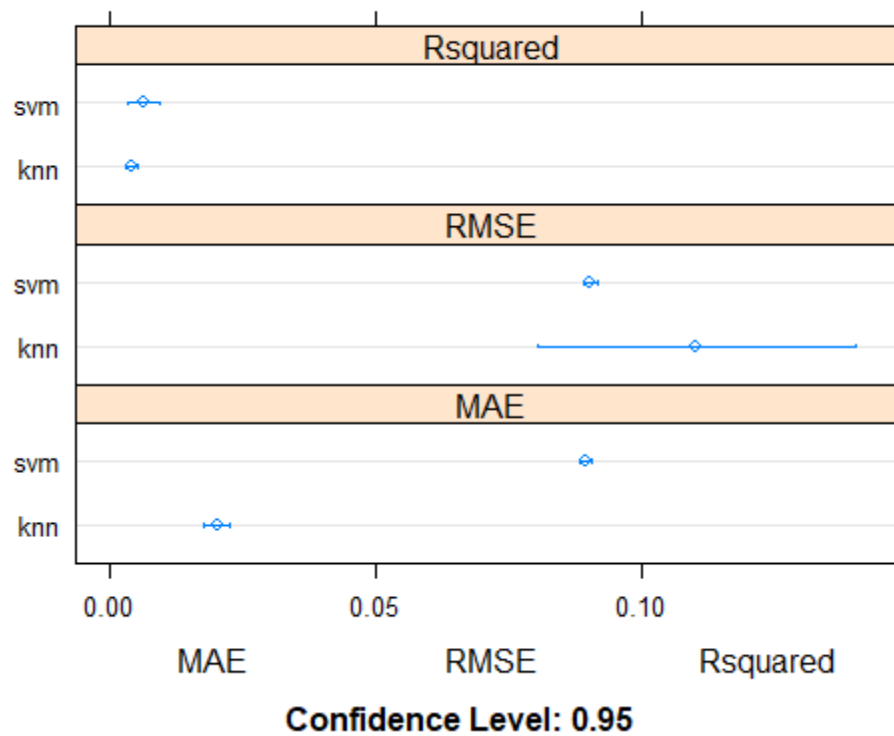
The boosting (also called Gradient Boosting Machine) ranks all of hundreds of independent variables and the following graph is showing the top ten important features. According to this algorithm, Game is the most important feature to predict and explain the PaidStarPerWatchedHour. Furthermore, PaidStarPerWatchedHour.1, Character Cont Openness, Role Strategist, Character Facet Cont Cooperation are other five highest important features.

To get a model assessment, then we tried to evaluate the model using testing set based on prediction ability. The MSE model is 0.00814 and the r square is 3.39%. the prediction of boosting model seems does not fit with the data well since the r square is very low,



c. KNN and SVM

In addition, we also applied KNN and SVM to get a solid decision. Both models have very low r square, since then both models are not suitable for the dataset.



5. Conclusion

- a. The dataset is very challenging since it provides hundreds of features.
- b. Although Female player is only less than 30%, the PaidStarPerWatchedHour Female player is much higher compare to male player.
- c. Another fact regarding female player, the games played both gender is more likely female player has upper hand rather than male players based on PaidStarPerWatchedHour.
- d. Paired variable for positive and negative correlations can be used for optimizing PaidStarPerWatchedHour.
- e. Several algorithms does not able to deal with hundreds of features, likely linear regression model with elimination method. Random forest also could not deal with variable consisted of more than 53 categories for instance Game.
- f. Random Forest, Gradient Boosting Machine, KNN, and SVM had been applied to detect the most important features on the dataset regarding PaidStarPerWatchedHour. According to model assessment, the highest r square model is Random Forest. This model is able to predict more than 50%.
- g. The most importante feature that can predict and optimize the PaidPerWatchedHour are PaidPerWatchedHour.1, Personal Value Facer Cont ConformityRules, UniversalismConcern, Character Cont Conscientiousness, character Facet Cont Modesty, Gender.