Prediction of League Level Based on StarCraft II Replay Data

Shenyi Pan

December 19, 2016

Abstract

This report studies the StarCraft II replay data collected by the Cognitive Science Lab at Simon Fraser University. The objective of this study is to predict the player league level based on the explanatory variables obtained from the replay games and examine which variables are important in distinguishing different league levels. it is found ordinal logistic regression provides relative high prediction accuracy rate on the league level along with high interpretability. The ordinal logistic regression model also suggests that during the same amount of time, players from higher-level leagues tend to have more actions, more perception action cycles, and more exploration and production activities in comparison to players from lower-level leagues.

1 Introduction

Electronic sports (eSports) are a form of competition that is facilitated by electronic systems, particularly video games. Typical eSports competitions are organized multiplayer competitions of video game between professional players. In recent years, the number of people watching or participating in eSports is rapidly increasing around the world. However, in order to become a top-tier eSports player, extensive training and commitment are required to develop the necessary skills and expertise. Professional eSports players usually perform consistently better than non-professional ones not only in terms of the overall game tactic and strategy, but also in terms of micromanagement skills such as how many actions they are able to execute per unit of time. In this project, we will analyze the replay data of StarCraft II, which is one of the most popular real-time strategy (RTS) game. The objective of this study is to predict the league level based on explanatory variables and investigate which variables are more important in characterizing StarCraft II professional players.

This report is organized as follows. Section 2 introduces the variables in the data set and the data cleaning procedure. Exploratory analysis on the variables and preliminary variable selection are performed in Section 3. Section 4 analyzes the replay data set with ordinal logistic regression.

Section 5 compares the prediction performance of the logistic regression model with the lowest AIC found in Section 4 with other machine learning methods. Finally, Section 6 concludes this report.

2 Data Summary and Data Cleaning

The data set used in this project was originally collected by the Cognitive Science Lab at Simon Fraser University (the data set is available on `http://summit.sfu.ca/item/13328`). This data set is an aggregate of the screen-fixations from screen movements of StarCraft II replay files of 3395 players across eight distinct levels of online competitive leagues ranging from novices to full-time professionals. The eight leagues are Bronze, Silver, Gold, Platinum, Diamond, Masters, Grandmaster, and Professional.

The response variable in this data set is *LeagueIndex*, where the eight leagues are coded as an ordinal variable ranging from 1 to 8 ranked from the lowest to the highest. In addition to the response variable, there are another 19 explanatory variables included in this data set. The detailed descriptions of all the variables can be found in Table 1.

Most of the variables in this data set are easy to understand based on their descriptions. However, the Perception Action Cycles (PACs) are a new terminology defined by the researchers at Simon Fraser University. In daily life, people use their attention to focus on objects they would like to interact with. In analogy, the vast majority of actions happen in StarCraft II are part of a Perception Action Cycle. A PAC basically consists of a shift of the screen to a new location for some time, followed by at least one action (typically 4-6 actions), and then a shift to some other location. In addition, time in this data set is recorded in terms of the timestamps in the StarCraft II replay file. When the game is played on the 'faster' mode in StarCraft II, 1 real-time second is equivalent to roughly 88.5 timestamps.

Before analyzing the data with statistical models, it is important to clean the data set and detect whether are any missing values or outliers. The summary statistics for each of the explanatory variables can be found is Table 2. From the summary statistics, it can be seen that the maximum value of the variable *TotalHours* is 1000000, while its minimum, first quartile, median, and third quartile are all smaller than 1000. Sorting the variable *TotalHours* further shows that the second largest value of this variable is 25000, which is also far smaller than 1000000 in magnitude. Therefore, we conclude that the value 1000000 is an outlier in *TotalHours*. Additionally, from Table 2, it can be seen that 55, 56, and 57 missing values for the variables *Age*, *HoursPerWeek*, and *HoursPerWeek*

Variable	Description	Type
LeagueIndex	Bronze, Silver, Gold, Platinum, Diamond, Master, GrandMaster, and Professional leagues coded 1-8	Ordinal
Age	Age of each player	Integer
HoursPerWeek	Reported hours spent playing per week	Integer
TotalHours	Reported total hours spent playing	Integer
APM	Action per minute	Continuous
SelectByHotkeys	Number of unit or building selections made using hotkeys per timestamp	Continuous
AssignToHotkeys	Number of units or buildings assigned to hotkeys per timestamp	Continuous
UniqueHotkeys	Number of unique hotkeys used per timestamp	Continuous
MinimapAttacks	Number of attack actions on minimap per timestamp	Continuous
MinimapRightClicks	number of right-clicks on minimap per timestamp	Continuous
NumberOfPACs	Number of PACs per timestamp	Continuous
GapBetweenPACs	Mean duration in milliseconds between PACs	Continuous
ActionLatency	Mean latency from the onset of PACs to their first action in milliseconds	Continuous
ActionsInPAC	Mean number of actions within each PAC	Continuous
TotalMapExplored	The number of 24x24 game coordinate grids viewed by the player per timestamp	Continuous
WorkersMade	Number of workers (SCVs, drones, or probes) trained per timestamp	Continuous
UniqueUnitsMade	Unique unites made per timestamp	Continuous
ComplexUnitsMade	Number of ghosts, infestors, and high templars trained per timestamp	Continuous
ComplexAbilitiesUsed	Abilities requiring specific targeting instructions used per timestamp	Continuous
MaxTimeStamp	Time stamp of game's last recorded event	Integer

Table 1: Descriptions and types for all the variables in the StarCraft II replay data set.

respectively. Although the number of missing values seems to be small in comparison to the total sample size 3395, it is worth noting that these three variables are missing for all the players from league 8, the highest league level. It is thus difficult to determine whether these three variables are important in distinguishing professionals from non-professionals with all the values missing for league 8 players. Therefore, we will exclude the variables *Age*, *HoursPerWeek*, and *HoursPerWeek* from our analysis.

From Table 1 we know that most of the variables in this data set such as *SelectByHotkeys*, *AssignToHotkeys*, *UniqueHotkeys*, *MinimapAttacks*, *MinimapRightClicks*, *NumberOfPACs*, *TotalMapExplored*, *WorkersMade*, *UniqueUnitsMade*, *ComplexUnitsMade*, and *ComplexAbilitiesUsed* are measured on a basis of timestamp. As a result, as is seen from Table **??**tab:summarystat, the magnitude of these variables are much smaller in comparison to the other variables. To make their magnitude similar to the other variables, we multiply each of the variables by $88.5 \times 60 = 5310$.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	No. of NA's
LeagueIndex	1.00	3.00	4.00	4.18	5.00	8.00	
Age	16.00	19.00	21.00	21.65	24.00	44.00	55
HoursPerWeek	0.00	8.00	12.00	15.91	20.00	168.00	56
HoursPerWeek	3	300	500	960.4	800	1000000	57
APM	22.06	79.90	108.01	117.05	142.79	389.83	
SelectByHotkeys	0	1.258e-03	2.500e-03	4.299e-03	5.1333e-03	4.309e-02	
AssignToHotkeys	0	2.042e-04	3.526e-04	3.736e-04	4.988e-04	1.752e-03	
UniqueHotkeys	0	3.275e-05	5.340e-05	5.873e-05	7.865e-05	3.376e-04	
MinimapAttacks	0	0	3.990e-05	9.831e-05	1.189e-04	3.019e-03	
MinimapRightClicks	0	1.401e-04	2.815e-04	3.874e-04	5.141e-04	4.041e-03	
NumberOfPACs	6.790e-04	2.754e-03	3.395e-03	3.463e-03	4.027e-03	7.971e-03	
GapBetweenPACs	6.67	28.96	36.72	40.36	48.29	237.14	
ActionLatency	24.09	50.45	60.93	63.74	73.68	176.37	
ActionsInPAC	2.04	4.27	5.10	5.273	6.03	18.56	
TotalMapExplored	9.130e-05	2.244e-04	2.695e-04	2.825e-04	3.253e-04	8.319e-04	
WorkersMade	7.70oe-05	6.830e-04	9.052e-04	1.032e-03	1.259e-03	5.149e-03	
UniqueUnitsMade	1.970e-05	6.780e-05	8.220e-05	8.455e-05	9.860e-05	2.019e-04	
ComplexUnitsMade	0	0	0	5.943e-05	8.555e-05	9.023e-04	
ComplexAbilitiesUsed	0	0	2.030e-05	1.419e-04	1.814e-04	3.084e-03	
MaxTimeStamp	25224	60090	81012	83598	102074	388032	

Table 2: Summary statistics (minimum, first quartile, median, mean, third quartile, maximum, and number of missing values) for all the explanatory variables in the StarCraft II replay data set.

After the transformation, these variables are measured per real-time minute when played on the faster mode in StarCraft II. On the contrary, the magnitude of the variable *MaxTimeStamp* is much larger in comparison to the other variables. We thus divide it by 5310 so that it measures the real-time minute of the game's length when played on the faster mode.

In addition, it is also important to take a look at the distribution of the response variable before performing any formal analysis. The bar plot of the counts for each league index in the StarCraft II replay data set is shown in the left-hand side of Figure 1. From the plot, it can be seen that the counts of league index 7 and 8 are fairly low compared with other indices. More specifically, there are only 35 observations of league 7, which is only around 1% of the total sample size. The insufficiency of observations makes it difficult to make predictions for leagues 7 and 8 and distinguish these two categories. To address this issue, we decide to merge leagues 7 and 8 and treat both of them as league 7. After merging these two leagues, the bar plot of the counts for each league index is shown in the right-hand side of Figure 1, where the counts for the last league is somewhat more comparable to other leagues.

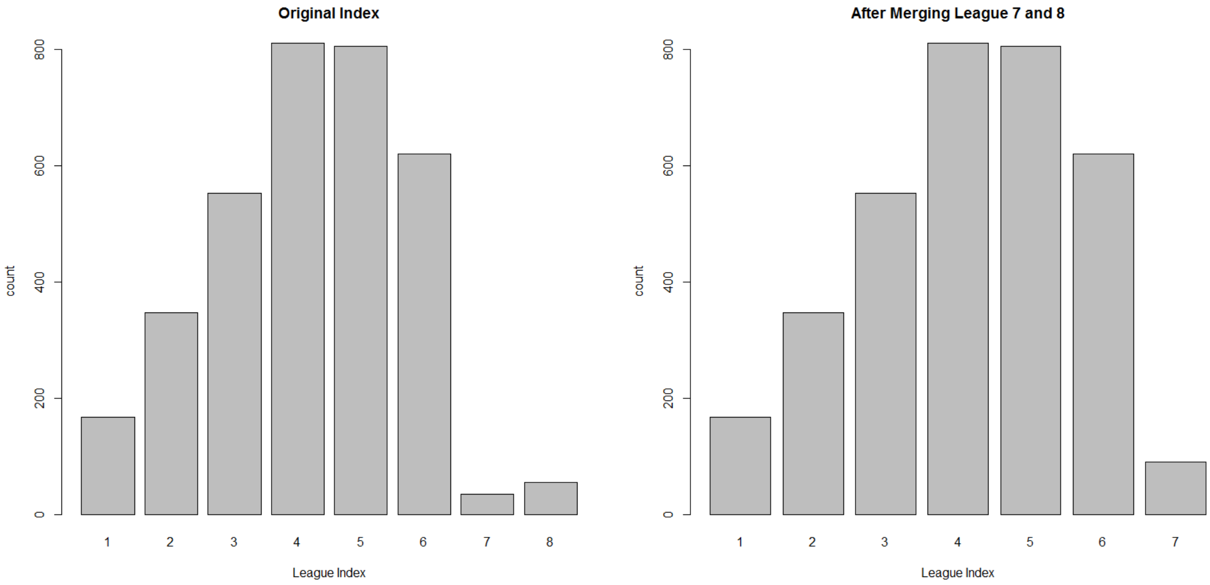

Figure 1: Left: The bar plot of the counts for each league index in the original StarCraft II replay data set. Right: The bar plot of the counts for each league index after merging leagues 7 and 8.

3 Exploratory Analysis and Variable Selection

In the previous section, we have introduced the StarCraft II replay data set that will be used in this project and performed some data preprocessing and cleaning work. In this section, we will conduct some explanatory analysis to examine which variables are more correlated with the response variable and select the more relevant variables to be included in our statistical model.

To check the dependence between all the variables, it is useful to compute the correlation matrix. However, since the number of variables is rather large in this data set, it might not be feasible to display the entire correlation matrix. Instead, the heat map of the correlation matrix can be found in Figure 2. Additionally, the correlation coefficient between the response variable *LeagueIndex* and all the explanatory variables is shown in Table 3. Figure 3 shows the box plots for the response variable and each of the explanatory variables. From the figures and the table, it can be seen that most of the explanatory variables are moderately correlated with *LeagueIndex* either positively or negatively. However, it is worth noting that the correlation between *MaxTimeStamp* and *LeagueIndex* is almost 0, which suggests the length of the game is not necessary correlated with the level of the league. The correlation between the response and the explanatory variables *ActionsInPAC*, *UniqueUnitsMade*, *ComplexUnitsMade* and *ComplexAbilityUsed* is also weak in comparison to other explanatory variables. Therefore, we will exclude the variables *MaxTimeStamp*, *ActionsInPAC*, *UniqueUnitsMade*, *ComplexUnitsMade* and *ComplexAbilityUsed*, and use the remaining 11 explanatory variables to build the ordinal logistic regression model in the next section.

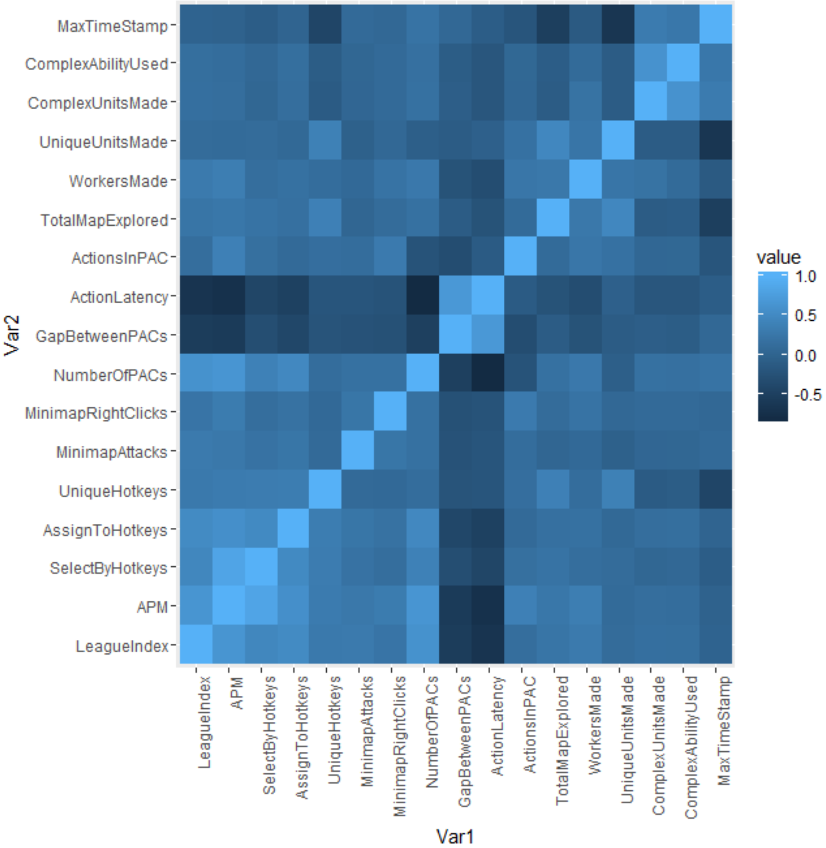

Figure 2: The heat map of the correlation matrix for the StarCraft II replay data set. Lighter color indicates stronger positive correlation while darker color indicates stronger negative correlation.

Variable	Correlation	Variable	Correlation
APM	0.6467	SelectByHotkeys	0.4680
AssignToHotkeys	0.5166	UniqueHotkeys	0.2875
MinimapAttacks	0.3018	MinimapRightClicks	0.2260
NumberOfPACs	0.6075	GapBetweenPACs	−0.5530
ActionLatency	−0.6742	ActionsInPAC	0.1436
TotalMapExplored	0.2334	WorkersMade	0.3019
UniqueUnitsMade	0.1135	ComplexUnitsMade	0.1557
ComplexAbilityUsed	0.1495	MaxTimeStamp	-0.0009

Table 3: Correlation coefficient between the response variable *LeagueIndex* and each of the explanatory variables.

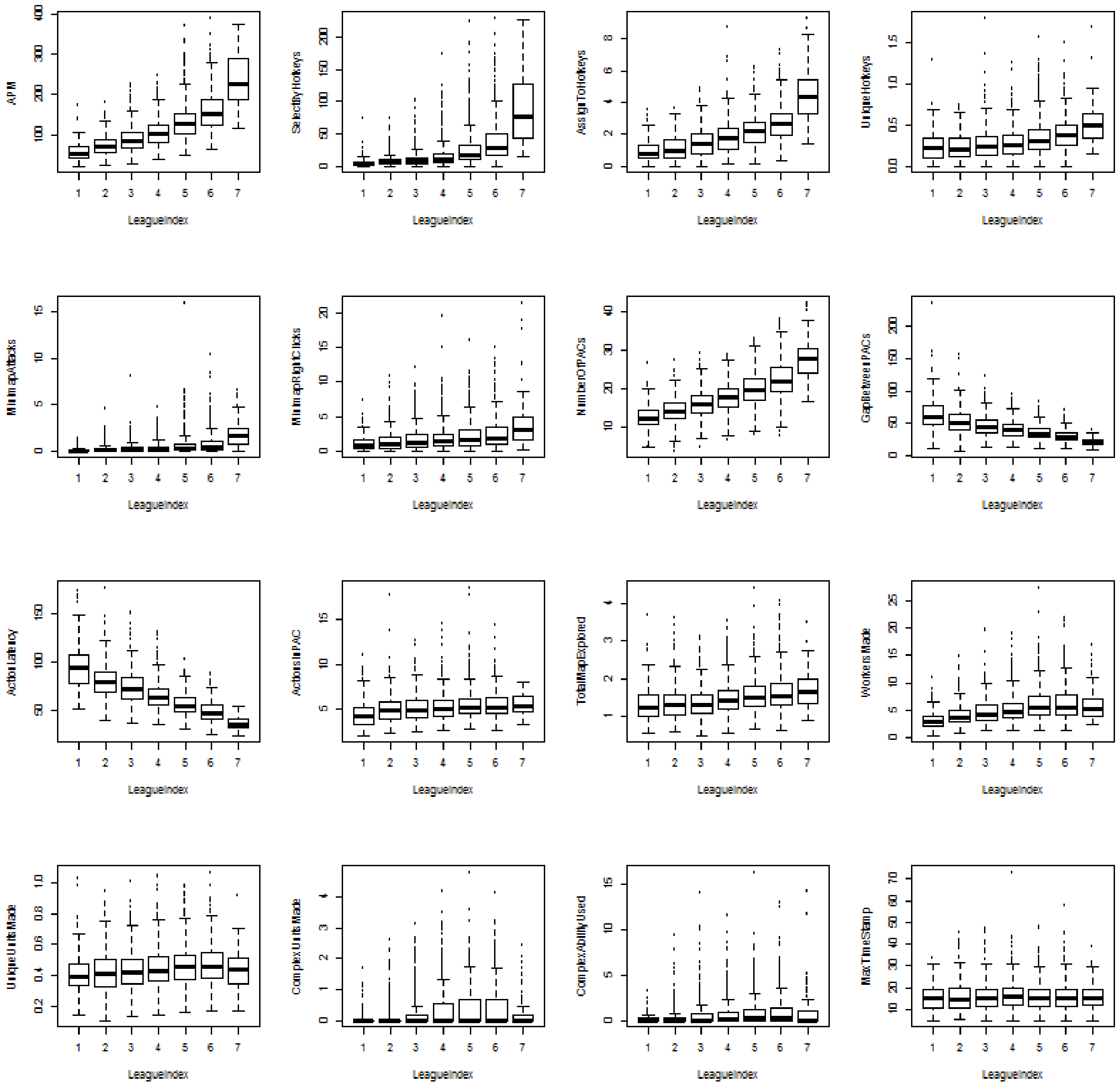

Figure 3: Box plots for the response variable *LeagueIndex* and each of the explanatory variables.

Another approaching to the selection of variables is to estimate the importance of variables by building a statistical model. For example, methods based on classification trees such as random forest use Gini impurity to report on variable importance. Larger drop of Gini impurity associated with one variable usually indicates higher importance. The ranked mean Gini impurity decrease for each of the explanatory variables in the StarCraft II replay data set can be found in Figure 4. From this figure, it can be seen that the variables *MaxTimeStamp*, *ActionsInPAC*, *UniqueUnitsMade*, *ComplexUnitsMade* and *ComplexAbilityUsed* all have relatively low importance based on fitting a random forest model (without parameter tuning via cross validation), which agrees with the previous result based on the correlation analysis.

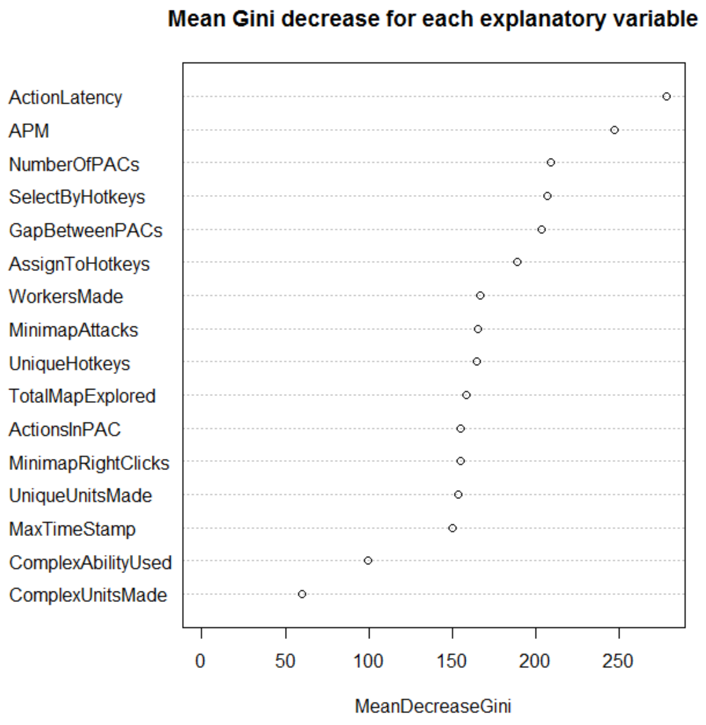

Figure 4: The mean Gini impurity decrease for each of the explanatory variables.

4 Analysis with Ordinal Logistic Regression

In the previous section, we have examined the relationship between the variables via exploratory analysis and removed some unimportant explanatory variables. In this section, we will build a prediction model for the league index of StarCraft II replay game based on ordinal logistic regression. In order to evaluate the prediction accuracy, we randomly select 1000 observations and treat them as the testing data set. The other 2395 observations are used as the training data set to fit the ordinal regression model.

Due to the large number of explanatory variables in the StarCraft II replay data set, we fit the ordinal logistic regression model assuming constant slopes for each category. Denote the data by (y_i, x_{i1}, x_{ip}), $i = 1, \ldots, n$, where y_i is the response, $\boldsymbol{x}_i$ are the explanatory variables, and n is the total number of observations in the training data set. There are in total 7 ordinal categories for the response *LeagueIndex*. Suppose Z is a latent variable that follows logistic distribution. Define the cut points $\eta_1 < \cdots < \eta_6$ as well as $\eta_0 = -\infty$ and $\eta_7 = \infty$. Then the ordinal logistic regression model is

$$\mathbb{P}(Y_i = k; \boldsymbol{x}_i) = \mathbb{P}(\eta_{k-1} - \beta_1 x_{i1} - \cdots - \beta_p x_{ip} < Z \leq \eta_k - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}),\ k = 1, \ldots, 7. \quad (1)$$

Since Z follows logistic distribution, this indicates that

$$\log \frac{\mathbb{P}(Y_i \leq k; \boldsymbol{x}_i)}{\mathbb{P}(Y_i > k; \boldsymbol{x}_i)} = \eta_k - \beta_1 x_{i1} - \cdots - \beta_p x_{ip},\ k = 1, \ldots, 6. \quad (2)$$

If we fit the ordinal logistic regression with constant slope for the StarCraft II replay data with *LeagueIndex* as the response variable and *APM*, *SelectByHotkeys*, *AssignToHotkeys*, *UniqueHotkeys*, *MinimapAttacks*, *MinimapRightClicks*, *NumberOfPACs*, *GapBetweenPACs*, *ActionLatency*, *TotalMapExplored*, and *WorkersMade* as the explanatory variables, the parameter estimates and the corresponding standard errors can be found in Table 4. The t-values in the table indicate that most of the variables are statistically significant except *MinimapRightClicks*. Stepwise model selection based on AIC eliminates the variable *MinimapRightClicks*, and the parameter estimates obtained from the reduced model are very similar to those in Table 4. The estimates for the parameters and the corresponding standard errors can be found in Table 5. We will denote this reduced model with the lowest AIC value by model A.

To interpret the results of Table 5, it can be seen that compared with players from lower-level leagues, during the same amount of time players from higher-level leagues tend to have more actions (e.g., selection by hotkeys, assignment to hotkeys, use of unique hotkeys, attacks on minimap), more perception action cycles, and more exploration and production activities (e.g., total map explored, workers made), but less latency between actions and gao between PACs. Nevertheless, it is curious that the explanatory variable *APM* is only marginally significant while during exploratory analysis it has the highest correlation with the response and it is also the second most important variable suggested by random forest model. Therefore, only including main effect terms might not capture all the information contained in the explanatory variables. The model fitting result could potentially be better if we also include the quadratic effects. If we fit the ordinal logistic regression model with both the main effects and the quadratic effects, stepwise model selection based on AIC eliminates the main effect of *MinimapRightClicks*, and includes the quadratic terms APM^2, $MinimapAttacks^2$, $NumberOfPACs^2$, $ActionLatency^2$, $TotalMapExplored^2$, and $WorkersMade^2$. The parameters have similar estimates and interpretations as in Table 4 and Table 5. We will denote this reduced model

Variable	Estimate	SE	t-value
Cut point 1|2	-2.9357	0.6492	-4.522
Cut point 2|3	-1.3056	0.6333	-2.062
Cut point 3|4	0.2228	0.6260	0.356
Cut point 4|5	1.8762	0.6263	2.996
Cut point 5|6	3.7364	0.6358	5.876
Cut point 6|7	7.2042	0.6845	10.524
APM	0.0046	0.0026	1.793
SelectByHotkeys	0.0111	0.0035	3.188
AssignToHotkeys	0.3121	0.0433	7.202
UniqueHotkeys	0.9886	0.2281	4.333
MinimapAttacks	0.3961	0.0497	7.965
MinimapRightClicks	-0.0057	0.0217	-0.263
NumberOfPACs	0.1031	0.0144	7.171
GapBetweenPACs	-0.0238	0.0035	-6.891
ActionLatency	-0.0299	0.0050	-5.931
TotalMapExplored	0.2233	0.0995	2.243
WorkersMade	0.0479	0.0159	3.010

Table 4: The estimate, standard deviation, and t-value for each parameter when fitting ordinal logistic regression with main effects only.

by model B.

After fitting these ordinal logistic regression models, we can compare the prediction accuracy of the two models A and B based on the testing data set. The table of predicted league index versus the true league index for each of these two models can be found in Table 6 and 7 respectively. In addition, some summary statistics that compare the prediction accuracy of these two models are listed in Table 8. Note that for both models most of the misclassifications occur when the predicted league indices are only one or two levels away from the observed ones. In this case, the weighted Cohen's kappa with absolute index difference as the weights might be a better choice to measure the agreement between the predicted and the observed league indices than the percentage agreement. From Table 8, it can be seen that the prediction performance of these two models are very close. For both models, the weighted Cohen's kappas are around 0.5, indicating moderate agreement between the predicted and the observed league indices. Considering the AIC and the overall prediction performance, we conclude that model B, the ordinal logistic regression model with both the main effects and the quadratic effects, is slightly superior to model A which only includes main effects.

Variable	Estimate	SE	t-value
Cut point 1|2	-2.9501	0.6469	-4.560
Cut point 2|3	-1.3198	0.6311	-2.091
Cut point 3|4	0.2088	0.6238	0.335
Cut point 4|5	1.8620	0.6240	2.984
Cut point 5|6	3.7217	0.6334	5.876
Cut point 6|7	7.1901	0.6825	10.536
APM	0.0044	0.0024	1.802
SelectByHotkeys	0.0113	0.0034	3.382
AssignToHotkeys	0.3118	0.0433	7.199
UniqueHotkeys	0.9880	0.2281	4.330
MinimapAttacks	0.3944	0.0493	8.003
NumberOfPACs	0.1033	0.0143	7.203
GapBetweenPACs	-0.0238	0.0035	-6.887
ActionLatency	-0.0301	0.0050	-5.979
TotalMapExplored	0.2220	0.0994	2.233
WorkersMade	0.0478	0.0159	3.003

Table 5: The estimate, standard deviation, and t-value for each parameter when fitting the reduced ordinal logistic regression with main effects only.

	Observed League Index						
Predicted League Index	1	2	3	4	5	6	7
1	11	17	3	2	0	0	0
2	11	14	15	9	2	0	0
3	10	47	52	46	12	1	0
4	4	25	62	116	77	17	0
5	0	5	13	75	107	74	0
6	0	0	2	9	50	88	11
7	0	0	0	0	0	3	10

Table 6: The table of predicted league index versus the true league index for model A.

	Observed League Index						
Predicted League Index	1	2	3	4	5	6	7
1	10	14	2	2	0	0	0
2	12	20	18	9	2	0	0
3	10	43	54	45	11	1	0
4	4	25	57	121	80	17	0
5	0	6	14	71	102	70	0
6	0	0	2	9	53	93	11
7	0	0	0	0	0	2	10

Table 7: The table of predicted league index versus the true league index for model B.

Model	Percentage agreement	Weighted Cohen's kappa	AIC
A	39.8%	0.513	6463.34
B	41.0%	0.521	6420.49

Table 8: Percentage agreement, weighted Cohen's kappa with absolute distance, as well as AIC value for ordinal logistic regression models A and B.

5 Performance Comparison with Other Classification Methods

In this previous section, we have analyzed the StarCraft II replay data set with ordinal logistic regression models and evaluated the prediction accuracy of these models. In this section, we will use other machine learning methods to predict the league index and compare their performance with the logistic regression models.

The machine learning methods that we will consider in this section include eXtreme Gradient Boosting (XGBoost), Stochastic Gradient Boosting Machine (GBM), Regularized Random Forest (RRF), and Support Vector Machines with Polynomial Kernel (SVMPoly). Although all these methods support both classification on a categorical variable and regression on a continuous variable, none of them are designed for classification problem on a ordinal variable. Considering the ordinal nature of the league index of the StarCraft II replay data set, we decide to fit a regression model on the training set with these four methods and then predict on the testing set. After obtaining the predictions, we will round them to the nearest integers and treat the rounded results as the predicted league indices. Measures of the prediction accuracy of these four methods on the testing data set are listed in Table 9. From the table, it can be seen that the reduced logistic regression model B with both the main effects and the quadratic effects have similar percentage agreement and weighted Cohen's kappa with these machine learning methods. Considering that ordinal logistic regression models are much more interpretable than the other machine learning methods, we conclude that the model B is a reasonable model for predicting the league index based on the StarCraft II replay data.

Model	Percentage agreement	Weighted Cohen's kappa
XGBoost	40.8%	0.502
GBM	40.2%	0.508
RRF	40.5%	0.492
SVMPoly	41.1%	0.507

Table 9: Percentage agreement and weighted Cohen's kappa with absolute distance for the four machine learning methods.

6 Conclusion

In this project, we have studied the StarCraft II replay data set and identified the best ordinal logistic regression model to predict the player's league index based on explanatory variables collected from the replay games. The analysis with ordinal logistic regression indicates there are a number of variables that are important in characterizing different leagues. In general, during the same amount of time, players from higher-level leagues tend to have more actions, more perception action cycles, and more exploration and production activities. Additionally, when compared with other machine learning methods, ordinal logistic regression provides similar prediction accuracy rate and higher inerpretability. Therefore, ordinal logistic regression is a reasonable prediction model for the StarCraft II replay data set.