STAT 311 Regression Analysis, Fall 2025

League of Legends Snowballing Analysis

November 14, 2025

Nicole Quentin & Charles Sandahl

Introduction

We would like to explore how snowballing (gaining an advantage over the enemy team and continuing to build upon that lead until the game ends) in League of Legends eSports beginning at 10 minutes. We want to explore variables that affect how much gold a team has 20 minutes into a match to see what areas are important to focus on.

Securing an early-game advantage increases a team's chances of winning the match. The amount of gold a team has 20 minutes into a game is a good metric for early-game snowballing, because teams that are behind will generally have less gold compared to if they were ahead. The variables we chose reflect a cumulative advantage that a team can get early in a game by performing well in the independent variables chosen.

We'd like to investigate what conditions can maximize how much gold a team has 20 minutes. We started by finding a dataset that as many variables as we could that would be relevant to our ability to predict snowballing. Then, we filtered down our dataset to find the variables that are the most relevant by looking for complete team data. After filtering, we found variables mattered the most when it came to predicting which team was in the lead for amount of gold earned at 20 minutes.

Data Description

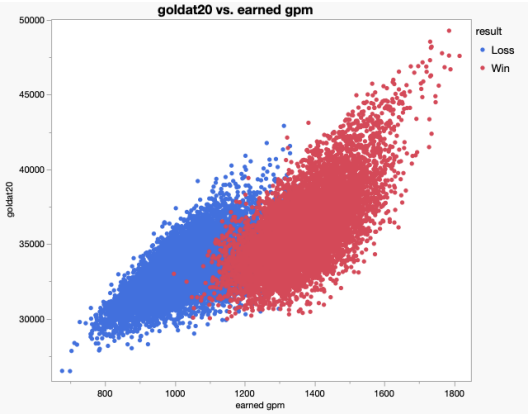
Our dataset comes from Oracle's Elixir, a website that publishes in-depth analytics for League of Legends. Oracle's Elixir collects data from multiple sources, including Riot Games' API, its official eSports website, and Leaguepedia, then automatically publishes the collated data on Google Drive (Oracle's Elixir).

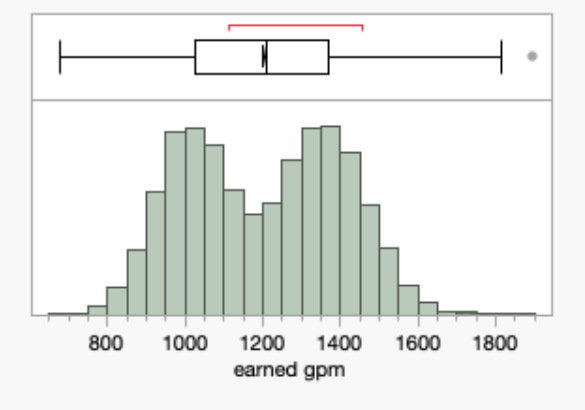
Oracle's Elixir's dataset could not exist without over 100,000 League of Legends eSports matches played in 2025 alone. Each match had different statistics and outcomes, and Oracle's Elixir automatically updated its dataset daily to account for each game. However, this leads to a massive dataset that has many independent variables and entries that are not useful to us. To fix this, we wrote a Python script that generates a filtered dataset that better suits our modeling. The filtered dataset only contains the columns with the variables we thought could be relevant for our model.

Exploratory Data Analysis

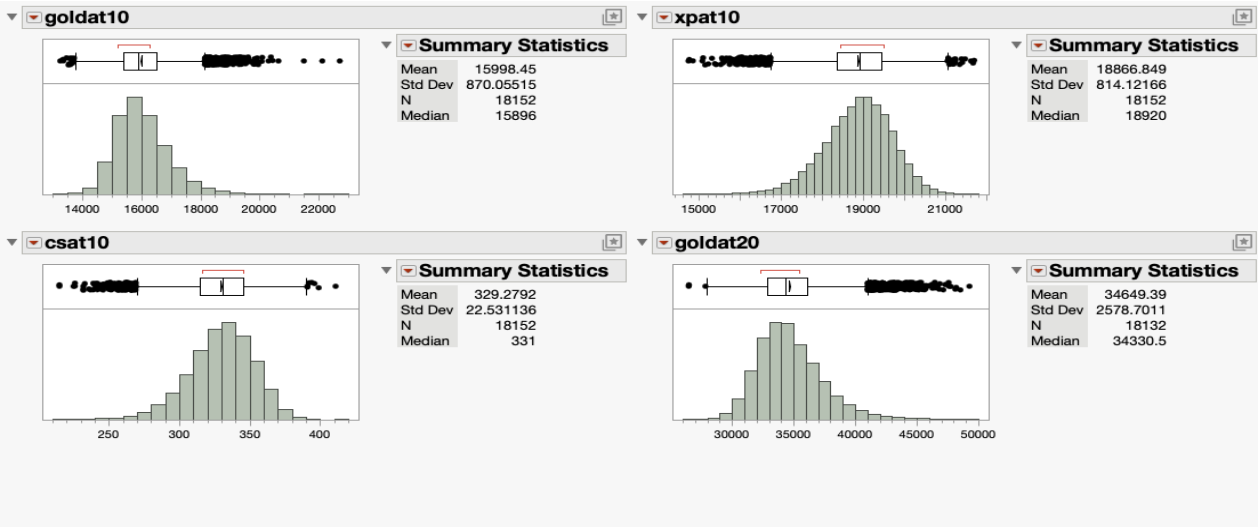
We originally intended to use the following variables as predictors because we thought they may be relevant to our research question: kills, deaths, assists, team KPM (game total kills/min.), CKPM (game total kills/min.), GSPD (game total gold spend percentage difference), DPM (game total damage/min.), CSPM (game total minions killed/min.), earned GPM (game total gold/min.), VSPM (game total vision score/min.), goldat10 (gold accumulated at 10 min. into the game), xpat10 (experience at 10 min. into the game), csat10 (minions killed at 10 min. into the game), and goldat20 (amount of gold at 20).

We noticed that the distribution of most of these variables were roughly normal, except earned GPM, which was bimodal. We predicted that earned GPM has a bimodal distribution because snowballing teams, who are more likely to win, earn more gold compared to the opposing team, who are more likely to lose. This hypothesis is supported by the following graphs:





This made us rethink the variables we had initially chosen. Many of the variables we had initially chosen were measured throughout the entire game, rather than the 10-minute to 20-minute window we want our model to represent. Additionally, removing these variables would make our model more parsimonious. After filtering for variables measured within the 20-minute time frame, the only continuous variables that remain are goldat10, xpat10, csat10, and goldat20.



Each variable follows a roughly normal distribution. Gold at 10 and Gold at 20 are slightly skewed to the right, while XP at 10 and CS at 10 are slightly skewed to the left. Gold at 10 and Gold at 20's means are slightly higher than their median, which shows how some teams get a gold lead early on. Gold at 20's standard deviation ranges from 870.05515 to 2578.7011,

implying that the game gets more spread out as the game progresses, which fits with our question. XP at 10 and CS at 10's mean is nearly equal to the median, showing that it is difficult to get a wide XP or CS lead this early in the game.

We noticed that the boxplots included with each variable's histogram showed many outliers were present in the data. When we fit the models, we will check if they are truly outliers or influential points.

Finally, for model construction, we have chosen these variables:

$$E(y) = goldat20, \quad x_1 = goldat10, \quad x_2 = xpat10, \quad x_3 = csat10$$

$$x_4 = \begin{cases} 1 = team\ took\ first\ tower \\ 0 = if\ not \end{cases}, \quad x_5 = \begin{cases} 1 = team\ took\ first\ dragon \\ 0 = if\ not \end{cases},$$

$$x_6 = \begin{cases} 1 = team\ took\ first\ herald \\ 0 = if\ not \end{cases}, \quad x_7 = \begin{cases} 1 = team\ got\ first\ blood \\ 0 = if\ not \end{cases}$$

Methods

Correlations

	goldat10	xpat10	csat10	goldat20	firsttower	firstdragon	firstherald	firstblood
goldat10	1.0000	0.3666	0.1557	0.6597	0.3306	0.0513	0.2607	0.2941
xpat10	0.3666	1.0000	0.7602	0.2868	0.2332	0.0695	0.1807	0.1214
csat10	0.1557	0.7602	1.0000	0.1560	0.2070	0.0433	0.1494	0.0562
goldat20	0.6597	0.2868	0.1560	1.0000	0.4706	0.0842	0.4041	0.2126
firsttower	0.3306	0.2332	0.2070	0.4706	1.0000	0.0465	0.2952	0.1421
firstdragon	0.0513	0.0695	0.0433	0.0842	0.0465	1.0000	0.0259	0.0632
firstherald	0.2607	0.1807	0.1494	0.4041	0.2952	0.0259	1.0000	0.1266
firstblood	0.2941	0.1214	0.0562	0.2126	0.1421	0.0632	0.1266	1.0000

There are 20 missing values. The correlations are estimated by Pairwise method.

Before running stepwise regression, we checked to see if there are any redundant variables by checking the multivariate correlations. In this table, we noticed that xpat10 and csat10 show multicollinearity with their multivariate values above 0.7, which means that one of

them is likely redundant. To determine which variable to keep, we calculated their R^2 values and chose the one that is larger:

$$\text{csat10 } R^2 = 0.1557^2 = 0.0\overline{24} < 0.13439 = 0.3666^2 = \text{xpat10 } R^2$$

Since we have determined that csat10 is redundant, we will use xpat10, goldat10, firsttower, firstdragon, firstherald and firstblood for stepwise regression.

Before running stepwise regression, it is helpful to run a validation step. We did this with training = 0.70, validation = 0.2 and test = 0.1 and training = 0.6, validation = 0.2 and test = 0.2 and then ran our stepwise regression using this as validation. The following table contains the top 5 models, sorted by C_p:

Top 5 models sorted by C_p using training = 0.7, validation = 0.2, test = 0.1

Model	R^2	AICc	BIC	C_p	Valid R^2	Test R^2
$x_1, x_2, x_4, x_5, x_6, x_7$	0.5374	225566.3095	225625.8892	7.0000	0.5522	0.5597
x_1, x_2, x_4, x_5, x_6	0.5374	225564.3098	225616.4432	5.0029	0.5522	0.5597
x_1, x_4, x_5, x_6, x_7	0.5374	225564.3731	225616.5065	5.0662	0.5522	0.5597
x_1, x_4, x_5, x_6	0.5374	225562.3738	225607.0605	3.0691	0.5522	0.5597
x_1, x_2, x_4, x_5, x_6	0.5360	225600.3747	225652.5081	41.0992	0.5497	0.5564

Top 5 models sorted by C_p using training = 0.6,validation = 0.2,test = 0.2

Model	R^2	AICc	BIC	C_p	Valid R^2	Test R^2
x_1, x_2, x_4, x_5, x_6	0.5440	193239.9930	193298.3350	7.0000	0.5194	0.5600
x_1, x_4, x_5, x_6	0.5440	193237.9906	193289.0412	5.0005	0.5194	0.5600

$x_1, x_2, x_4, x_5, x_6, x_7$	0.5439	193240.6756	193291.7261	7.6841	0.5198	0.5605
x_1, x_4, x_5, x_6, x_7	0.5439	193238.6730	193282.4317	5.6841	0.5198	0.5605
x_1, x_2, x_4, x_6	0.5424	193276.0398	193327.0904	43.0918	0.5178	0.5575

The variables that show up in nearly all models between the validation data are goldat10, firsttower, firstdragon, and firstherald. There is not much variance between the validation data, but when there is, we noticed that introducing xpat10 and firstblood increases the BIC and $C_p$values with minimal effect to the R^2 values. The model with the lowest BIC and $C_p$values include goldat10, xpat10, firsttower, firstdragon, and firstherald, and the second model includes only goldat10, firsttower, firstdragon, and first herald.

First Proposed Model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$

Second Proposed Model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$

To test which of the two models is best, we ran a 5-Fold Cross-Validation. When running forwards and backwards, it removed xpat10 which has a p-value of 0.9232. The R^2 K-fold value is 0.5426, which is close the R^2 values for both sets of training data. Finally, the $C_p$value is 4.0093, which is lower than the number of parameters in our dataset plus one $(k + 1)$, indicating that our second model is likely the most unbiased.

Model 1

Proposed model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$,

where $E(y)$ = expected goldat20, $x_1 = goldat10$, $x_2 = \begin{cases} 1 = team\ took\ first\ tower \\ 0 = if\ not \end{cases}$,

$x_3 = \begin{cases} 1 = team\ took\ first\ dragon \\ 0 = if\ not \end{cases}$, $x_4 = \begin{cases} 1 = team\ took\ first\ herald \\ 0 = if\ not \end{cases}$

Fitted model: $\hat{y} = 8265.063 + 1.573x_1 + 1220.052x_2 + 218.549x_3 + 1010.881x_4$

Global F-test:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	6.5445e+10	1.636e+10	5380.514
Error	18127	5.5121e+10	3040821.3	Prob > F
C. Total	18131	1.2057e+11		<.0001*

H_0: $All\ \beta_i = 0$. H_1: $At\ least\ one\ \beta_i \neq 0$. $(i = 1, 2, 3, 4)$

p-value < 0.0001 $\alpha = 0.05$ p-value $< \alpha$

At a 5% significance level, we reject the null hypothesis. There is sufficient evidence to indicate that Model 3 is statistically useful for predicting goldat20.

R_a^2 and MSE:

Summary of Fit

RSquare	0.542814
RSquare Adj	0.542713
Root Mean Square Error	1743.795
Mean of Response	34649.39
Observations (or Sum Wgts)	18132

$R_a^2 = 0.542713$ $s = 1743.795$ $2s = 3487.59$

R_a^2 interpretation: Approximately 54.27% of variation in amount of gold at 20 minutes can be explained by the independent variables, adjusted for sample size and the number of independent variables in the model.

MSE interpretation: We expect approximately 95% of the observed amount of gold at 20 minutes to fall within 3,487.59 gold of the model's predictions.

Cross-validation: When this model is cross-validated with 70% of data allocated for training and 30% for validation, the model's overall R^2 value increases from 0.5428 to 0.5374. Additionally, there isn't a large difference in R^2 values for the training set and validation set (0.5522 vs. 0.5597).

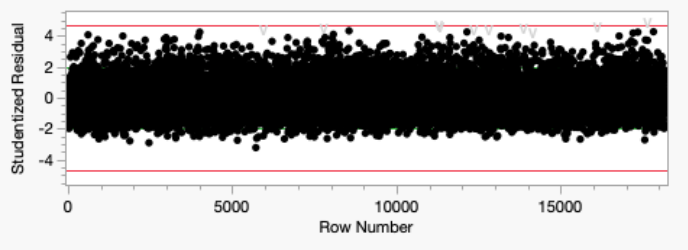
All Possible Models

Ordered up to best 4 models up to 6 terms per model.

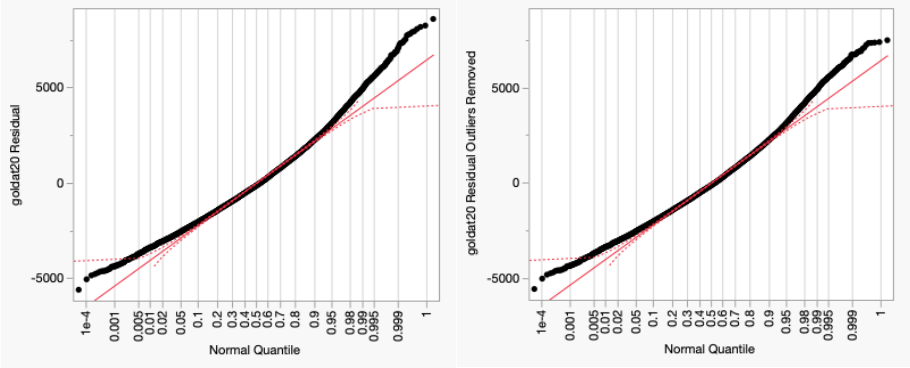
Model	Number	RSquare	RMSE	AICc	BIC	Cp	Validation RSquare	Test RSquare
goldat10,firsttower,firstdragon,firstherald	4	0.5374	1746.54	225562	225607	3.0691	0.5522	0.5597

Checking Model Assumptions

Outliers: The Studentized residual graph below shows the points that are closest to the 95% cutoff as 'v'. After removing those points, the Q-Q plot doesn't change much and still has a similar shape. The R_a^2 value also only increases from 0.5427 to 0.5431, so we checked the point with the highest Cook's distance next. The row with the highest Cook's Distance is 0.0045, which is much less than 1, so we will keep the outliers in our model.

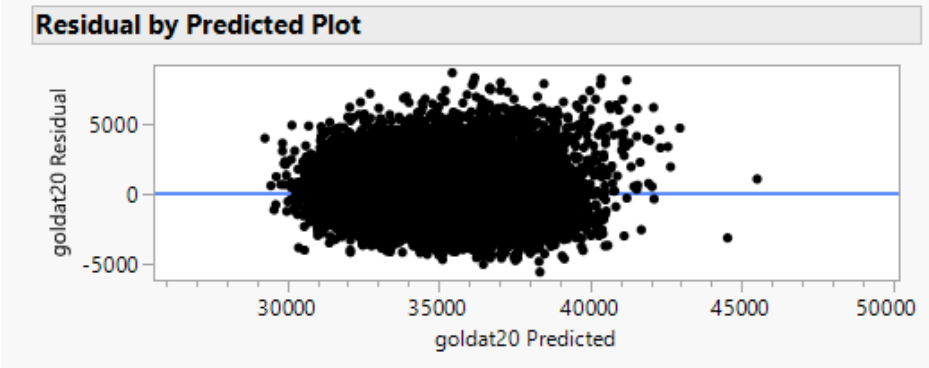


Normality Assumption: Model 3 with/without outliers roughly fit the straight line in the Normal Quantile plot. We believe the Normality Assumption is satisfied.



Lack of Fit: We don't notice any distinct trends in the plot below. We believe the Lack of Fit assumption is satisfied.

Unequal Variances: We do not have enough evidence to say the Unequal Variances assumption is satisfied. There is no distinct pattern in the Residual by Predicted plot. However, we attempted a square root transformation, $\sin^{-1}$, and ln, but there was no difference.



Model 2

Proposed model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4$

Fitted model: $\hat{y} = 45979.632 - 3.064x_1 + 0.0001x_1^2 + 1237.813x_2 + 211.402x_3 + 1019.799x_4$

Global F-test:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	6.5967e+10	1.319e+10	4379.968
Error	18126	5.4599e+10	3012196.7	Prob > F
C. Total	18131	1.2057e+11		<.0001*

$H_0: All\ \beta_i = 0.$ $H_1: At\ least\ one\ \beta_i \neq 0.$ $(i = 1, 2, 3, 4, 5)$

p-value < 0.0001 $\alpha = 0.05$ p-value $< \alpha$

At a 5% significance level, we reject the null hypothesis. There is sufficient evidence to conclude that the overall model is statistically useful for predicting goldat20.

T-test on Beta:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	45979.632	2876.196	15.99	<.0001*
goldat10	-3.06387	0.352618	-8.69	<.0001*
goldat10*goldat10	0.000142	1.079e-5	13.16	<.0001*
firsttower	1237.8132	28.09663	44.06	<.0001*
firstdragon	211.40195	25.82813	8.18	<.0001*
firstherald	1019.7993	27.43317	37.17	<.0001*

$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$ p-value < 0.0001 p-value $< \alpha$

Since the p-value is smaller than α, we reject the null hypothesis at a 5% significance level. There is sufficient evidence to conclude that $goldat10^2$ is useful for predicting goldat20.

R_a^2 and MSE:

Summary of Fit	
RSquare	0.547143
RSquare Adj	0.547018
Root Mean Square Error	1735.568
Mean of Response	34649.39
Observations (or Sum Wgts)	18132

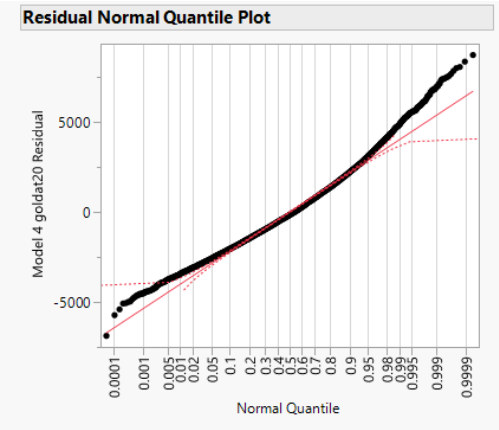
$R_a^2 = 0.547018$ $s = 1735.568$ $2s = 3471.136$

R_a^2 interpretation: Approximately 54.7% of variation in amount of gold at 20 minutes can be explained by the independent variables, adjusted for sample size and the number of independent variables in the model.

MSE interpretation: We expect approximately 95% of the observed amount of gold at 20 minutes to fall within 3,471.136 gold of the model's predictions.

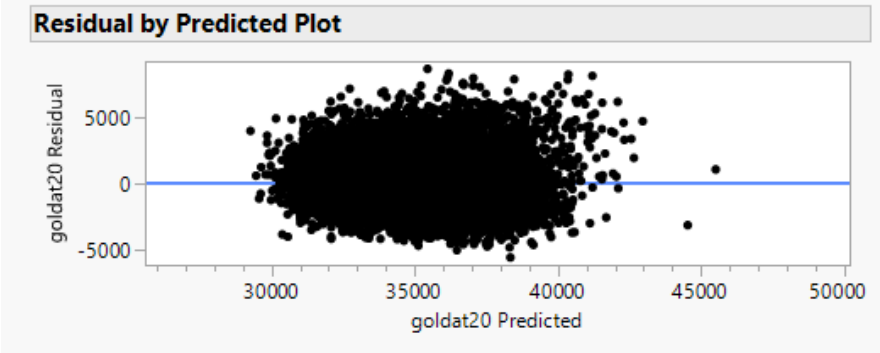
Checking Model Assumptions

Normality Assumption: Compared to Model 3, the Q-Q plot has become slightly more linear, especially in the highest areas. Since it roughly fits the straight line, we believe the normality assumption is satisfied:



Lack of Fit: There's no distinct curvilinear trend in the Residual vs. Predicted plot. We believe the Lack of Fit assumption is satisfied.

Unequal Variances: We do not have enough evidence to claim the Unequal Variances assumption is satisfied or not satisfied. There is no distinct pattern in the Residual by Predicted plot. However, we attempted a square root transformation, $\sin^{-1}$, and ln, but there was no difference.



Model 3

Proposed model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_3 + \beta_6 x_1 x_3 + \beta_7 x_4 + + \beta_8 x_1 x_4$

Fitted model: $\hat{y} = 36182.604 - 1.649 x_1 + 0.00009 x_1^2 - 887.349 x_2 + 0.133 x_1 x_2 - 630.419 x_3 + 0.052 x_1 x_3 - 2290.059 x_4 + 0.207 x_1 x_4$

Global F-test:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	6.6163e+10	8.2704e+9	2755.096
Error	18123	5.4403e+10	3001852	Prob > F
C. Total	18131	1.2057e+11		<.0001*

H_0: *All* $\beta_i = 0$. H_1: *At least one* $\beta_i \neq 0$. ($i = 1$ through 8)

p-value < 0.0001 $\alpha = 0.05$ p-value $< \alpha$

Due to the p-value being smaller than α, we reject H_0 at a 5% significance level. There is sufficient evidence to conclude that Model 5 is statistically useful for predicting goldat20.

Partial F-test:

Custom Test

Parameter				
Intercept	0	0	0	0
goldat10	0	0	0	0
goldat10*goldat10	1	0	0	0
firsttower	0	0	0	0
goldat10*firsttower	0	1	0	0
firstdragon	0	0	0	0
goldat10*firstdragon	0	0	1	0
firstherald	0	0	0	0
goldat10*firstherald	0	0	0	1
=	0	0	0	0
Value	0.0000918015	0.1330144034	0.0526157451	0.2070632827
Std Error	0.0000126435	0.0369817401	0.0300017371	0.0343409909
t Ratio	7.2607372471	3.5967589203	1.7537566208	6.0296245805
Prob>|t|	4.005793e-13	0.0003230638	0.0794891213	1.6752681e-9
SS	158252551.62	38833983.209	9232683.0518	109136450.58

Sum of Squares	718403195.38
Numerator DF	4
F Ratio	59.829997442
Prob > F	2.807835e-50

H_0: $All\ \beta_i = 0$ H_1: $At\ least\ one\ \beta_i \neq 0\ (i = 2, 4, 6, 8)$ p-value = 0+

p-value $< \alpha$

Since the p-value is smaller than α, we reject the null hypothesis. At a 5% significance level, there is sufficient evidence to indicate that $goldat10^2$, and the interactions between goldat10 and firsttower, goldat10 and firstdragon, and goldat10 and firstherald are statistically useful for predicting goldat20.

R_a^2 and MSE:

Summary of Fit	
RSquare	0.548772
RSquare Adj	0.548573
Root Mean Square Error	1732.585
Mean of Response	34649.39
Observations (or Sum Wgts)	18132

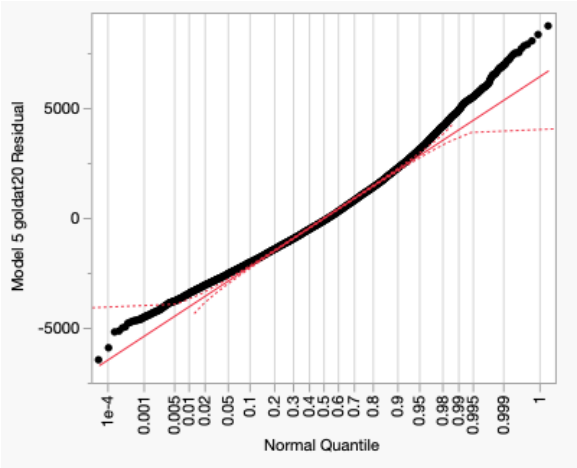
$R_a^2 = 0.548573$ $s = 1732.585$ $2s = 3465.17$

R_a^2 interpretation: Approximately 54.8% of variation in amount of gold at 20 minutes can be explained by the independent variables, adjusted for sample size and the number of independent variables in the model.

MSE interpretation: We expect approximately 95% of the observed amount of gold at 20 minutes to fall within 3,465.17 gold of the model's predictions.

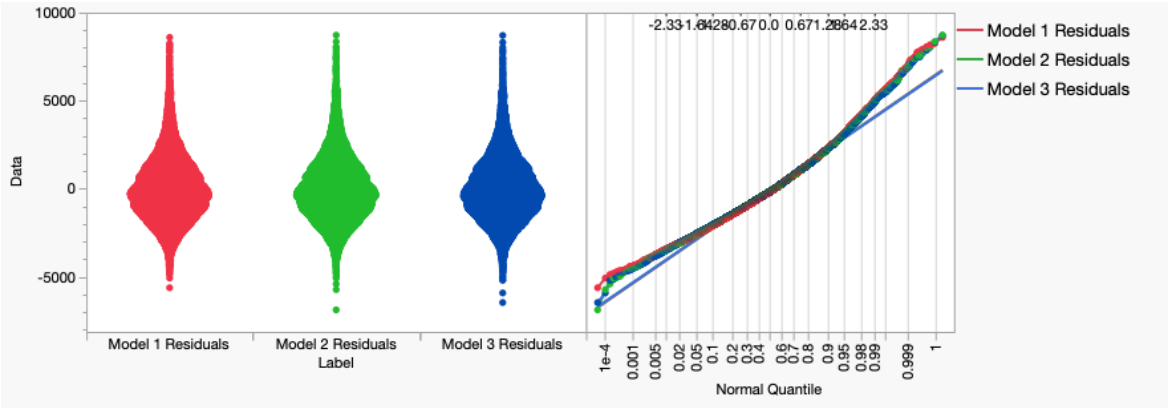
After adding interactions between variables, the Q-Q plot becomes slightly more linear like Model 4, especially in the highest areas. R_a^2 also slightly increased again to 0.5485. Despite the increase in R_a^2 and rejecting the null hypothesis for the Global F-test and Partial T-test, the model does not make sense practically. Destroying turrets grants extra gold, and objectives like

Dragon and Rift Herald both grant gold on takedown, so it does not make sense for their beta estimates to be negative. This coupled with adding 3 more variables makes it not very useful even with the adjusted R^2 value at its highest. Due to these reasons, we will not formally check Model 5's assumptions.

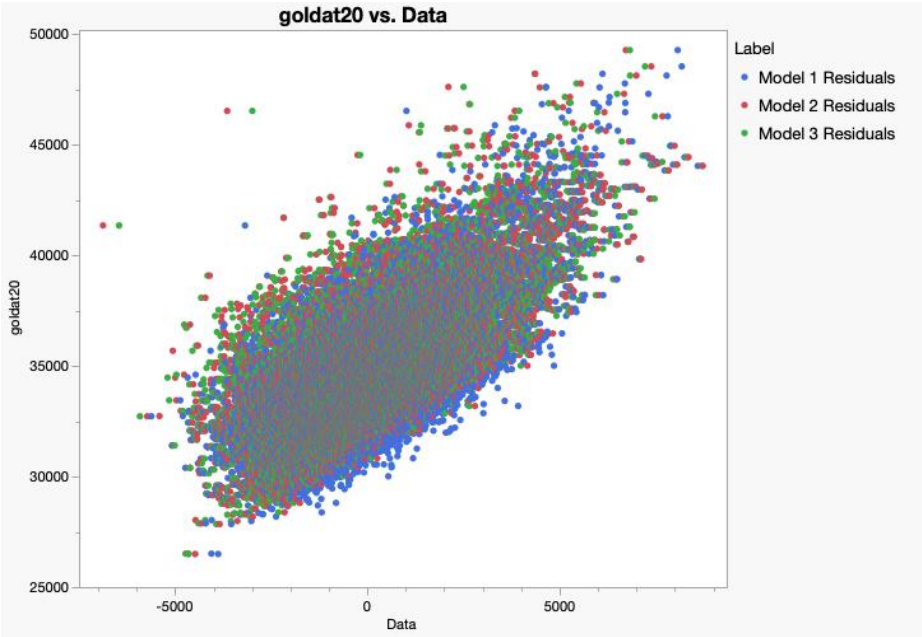


Results

For our findings, we chose to look at the parts of our methods that we checked for each model. We will compare Q-Q plots, Residuals, Mean Square Error, Adjusted R^2, and F-Score.



Our Q-Q and Violin plots show that the fit of the models is almost the same. They show that there is the same loose linear relationship to them. Using this as our only method to choose a model is inconclusive.



Next, we look at the residuals and notice that they also have the same shape to them. They do not show any obvious pattern that would indicate heteroscedasticity and nullify the assumption of unequal variance. There is enough evidence to suggest that the variance in all the models is equal, therefore, we will not use this as a method to choose our model.

Model	R_a^2	MSE	Global F-Score
Model 1	0.542713	1743.795	<0.0001
Model 2	0.547018	1735.568	<0.0001
Model 3	0.548573	1732.585	<0.0001

Finally, we look at variance of the models. According to their F-Scores, there is sufficient evidence to indicate that all three models are useful in predicting expected goldat20. Model 1 has the lowest R_a^2 and highest MSE among all three models, but the difference is very small. Model 1 is also the most parsimonious of the three models.

In Model 1, all independent variables have a positive effect on the amount of gold a team has at 20 minutes. This changed once we added interactions between variables. In Model 2, goldat20 decreases by 3.064 for every one-gold increase in goldat10. In Model 3, goldat20 decreases by 1.649 for every one-gold increase in goldat10, decreases by 887.349 if the team took the first tower, decreases by 630.419 if the team took the first dragon, and decreases by 2290.059 if the team took the first dragon. Due to Model 2 and 3 having negative beta estimates that do not make sense practically, we decided not to use them. These results also convinced us that there are no interactions between goldat10, and firsttower, firstdragon, and firstherald.

We believe that the best model out of all three that we trained is Model 1:

$$\hat{y} = 8265.063 + 1.573x_1 + 1220.052x_2 + 218.549x_3 + 1010.881x_4$$

It is a simple first order model which has the lowest variance, but the least number of variables. Adding a second order term and interactions does not improve R_a^2 enough for the increase in complexity to be worth it. We expect teams to have about 8265 gold at the 20-minute mark with an additional 1.5 gold for each gold a team has at 10 minutes. Our model shows that if the first tower, first dragon, and first herald are taken, the gold expected from completing those objectives would be about 1220, 218, and 1010 gold respectively. However, we do not believe that this is the best model that can be created with this dataset. If we used different variables, we may be able to increase R_a^2 past 54%.

Conclusions and Discussion

In this project, we explored how snowballing in League of Legends occurs before 20 minutes into a match. More specifically, we tried to see if we could accurately predict the amount of gold a team in League of Legends has at 20 minutes into a game, by using the amount of gold they have at 10 minutes, in addition to whether the team completed specific objectives: first tower, first dragon, and first herald.

The best model we could fit with the variables we chose explained about 54% of the variance in the model. Given the chaotic and noisy environment of an eSports game like League of Legends, this is an acceptable amount of variance, and the model could be a semi-accurate predictive model. We believe that we could possibly increase the model's accuracy further by including additional variables, but there are two concerns: we need to also be sure that additional variables are not measured over the course of the entire game rather than before 20 minutes, and we need to ensure that our future models do not become overly complex by adding too many variables.

References

Oracle's Elixir. *Frequently Asked Questions*. n.d. Document. 14 November 2025. <https://oracleselixir.com/faq>.

Oracle's Elixir. *Google Drive*. 14 January 2025. CSV. 5 November 2025. <https://drive.google.com/file/d/1v6LRphp2kYciU4SXp0PCjEMuev1bDejc/view?usp=sharing>.