KWALEE Data Analyst Test

Dataset owner: KWALEE
Solution provider: MANOJ ASHVIN JAYARAJ
Date: 07-January-2020

Introduction:

Kwalee's dataset contains information about a random selection of 5,000 players who installed one of their apps in the last ~30 days. The dataset contains only one file named as Level_progress.csv.

The columns in the level_progress.csv file are defined as follows.

- event_datetime: The date and time at which the event was received (Local time)
- player_id: A unique identifier for each player.
- level_number: The level number the event corresponds to.
- status: The outcome the event corresponds to. (start, fail, complete)
- session_id: A unique identifier for the session in which the event was produced.

*complete means the player was successful.

Task

Question - On which level are players most likely to fail?

Hint: You should consider the statistical significance of your answer carefully.

Solution Approach

Predicting the probability of success or failure in each level can help us to determine on which level a player is most likely to file. So during the data preparation phase the records with start status at a level will be skipped. To understand the statistical significance we start with a Generalized Linear Model with a Logit Link function because we only have two possible outcomes i.e. completed and failed.

Approach:

1. Understand the problem
2. Understand the Data and decide whether it's possible to answer the question
3. **Go/No Go - If YES in step 2** *then move to step 4* **else** *redefine the problemstatement*
4. Explanatory Data Analysis
5. Data Preparation
6. Use statistical modeling technique
7. Conclusion
8. Future works

1. Understand the Problem:

Kwalee wants to determine on which level a player is most likely to fail. To determine the level where players are most likely to fail, researcher need information of players who played different levels in the game. In reality not all gamers will succeed at every level and so the data should exhibit failure and success at different levels to answer the question.

2. Data Understanding:

Given dataset has nearly 55k entries corresponding to players starting a level and proceed to either complete the level in single attempt or take multiple attempts to complete i.e. failing multiple times and then completing it. The data has enough historical observations to make a prediction for future, which is important for a statistical model. Also the data has information of whether a person completed/failed a level in single or multiple attempts. This information can give more explanatory power to predict the failure probability. Also we have information on what day and time the player played the game which can add some more power to our prediction. Give the above information ***we have enough information to start a statistical model to predict the failure probability at different levels.***

3. Go/No Go: Based on the data understanding, we have enough evidence in the data to determine the level which has most likelihood of a failure. Hence we move to step 4 i.e. Explanatory Data Analysis.

4. Explanatory Data Analysis

A plot of no of entries corresponding to different status recorded (i.e. start a level, fail in a level, and complete a level) as in **Fig: 1** revels that there is **imbalance in the status recorded**. An imbalance dataset will not produce a good prediction because of not enough observation in some segments of prediction. This has to be taken into account to increase the no of samples using SMOTE technique. We will discuss more about SMOTE in modeling phase.

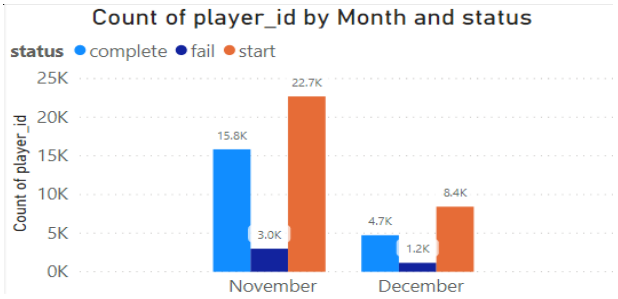


Fig: 1 No of entries in each status

The given data was a record of player activities over 30 day period. Plotting the status recorded over 30 days as in **Fig: 2** reveal the imbalance in status across every day. **If we are considering the temporal aspect from the dataset then the SMOTE technique should add a balanced entry in each day or else the statistical results will not be accurate.** But in this experiment we will ignore the temporal aspect of the data and try to model it without the temporal nature preserved in the dataset.

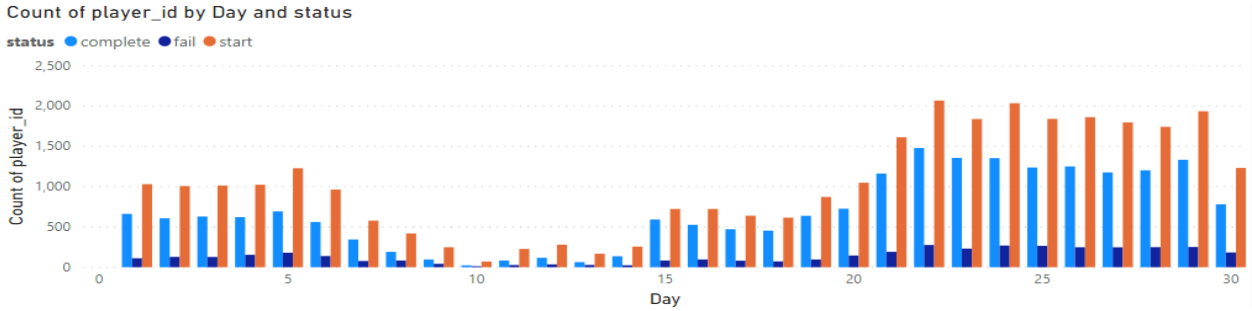


Fig: 2 Status recorded over a 30 day period

A plot of entries recorded in each level as in **Fig: 3** reveals that there are in total 49 levels covered by different players in the game. The skewed pattern reveals that not everyone has played till level 49 and the involvement in participation decreases as the level number increases. From **Fig: 2** we can also see that more people started playing the game towards the last 10 days of the month and that could also be a reason for not all player able to reach till level 49.

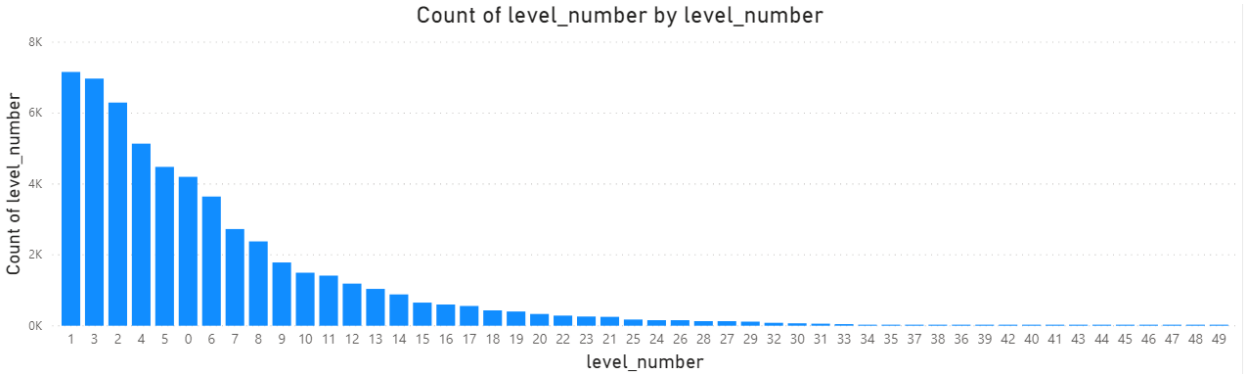


Fig: 3 No of entries recorded at each level

Fig: 4 reveals that people play the game round the clock, however majority of players play the game between 7am and 12pm and the peak is observed at 7 pm. No of players playing the game gradually decreases after 7pm. Since we don't have the age variable in our dataset it's hard to differentiate which age group plays the game during the day and which group plays in the evening and which group plays in the night.

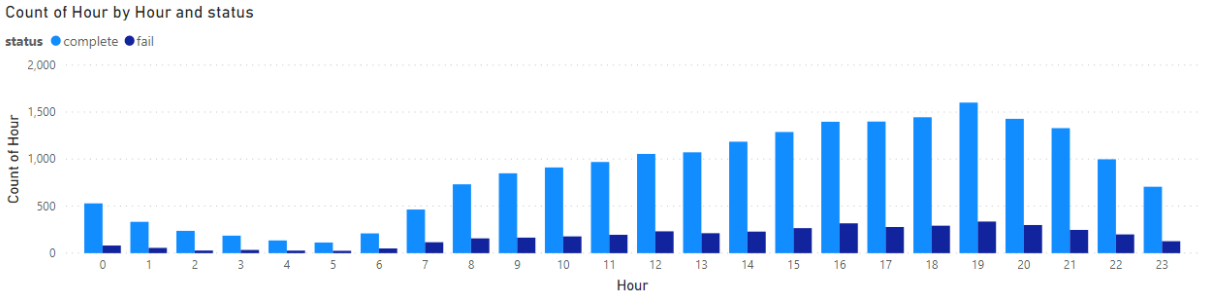


Fig: 4 Play time recorded in each day

The most important information of all is the number of players completing/failing at different levels in the game which is given in **Fig: 5.** Its visible that many people played the first 10 levels in the game and there after the number goes down which could be attributing to the observation noted in **Fig: 4**. The count of failure corresponds to the count of completing by some factor but the count of failure never exceeds the rate of completion in any of the 49 levels observed in the dataset. Based on the above exploration my human institution states that level no 3,5,6,7 can have a significance failure count compared to other levels. Whereas levels beyond 15 have less number of observations and it's difficult for a human to make a guess but the statistical models can explain it. It will be interesting to observe if a statistical model can find any levels beyond 15 to be significant in determining the failure. We will come back to discussion after observing the model results.

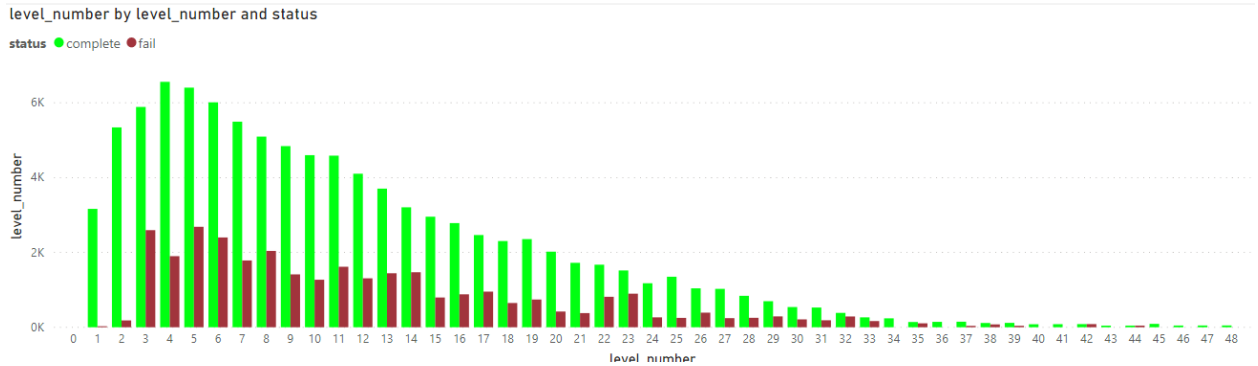


Fig: 5 Success and failure at all the levels registered in the dataset

Looking at individual player's data revealed some pattern in the data that requires cleaning/removal of such patterns, if not it will affect the statistical modeling results. **Fig: 6** show some entries corresponding to "*player_id": 00576ef3617e6bbbabfa0b1090e4122d,* where the player started level 7 on 3rd December 2020 and the session is marked as "*session_id": 21fd6a7db53a9220574563ac8a8c8b58* followed by two subsequent failures in the same session. Then the player started a new session on the next day and started the game twice in the same session without failing or completing level 7. As the data is sorted based in "event_date" it's strange to observe two start status in level 7th on December 4th without a

completion or failure status. Such entries are inconsistent and can affect the statistical modeling results and so it should be removed from modeling.

player_id	session_id	status	level_number	event_datetime
00576ef3617e6bbbabfa0b1090e4122d	21fd6a7db53a9220574563ac8a8c8b58	complete	5	12/3/2020 12:07:33 PM
00576ef3617e6bbbabfa0b1090e4122d	21fd6a7db53a9220574563ac8a8c8b58	start	6	12/3/2020 12:07:58 PM
00576ef3617e6bbbabfa0b1090e4122d	21fd6a7db53a9220574563ac8a8c8b58	complete	6	12/3/2020 12:11:18 PM
00576ef3617e6bbbabfa0b1090e4122d	21fd6a7db53a9220574563ac8a8c8b58	start	7	12/3/2020 12:12:12 PM
00576ef3617e6bbbabfa0b1090e4122d	21fd6a7db53a9220574563ac8a8c8b58	fail	7	12/3/2020 12:16:10 PM
00576ef3617e6bbbabfa0b1090e4122d	21fd6a7db53a9220574563ac8a8c8b58	fail	7	12/4/2020 2:18:10 PM
00576ef3617e6bbbabfa0b1090e4122d	759dc23452c8a3d3e64e534638e17b7f	start	7	12/4/2020 2:18:54 PM
00576ef3617e6bbbabfa0b1090e4122d	759dc23452c8a3d3e64e534638e17b7f	start	7	12/4/2020 2:22:04 PM
00576ef3617e6bbbabfa0b1090e4122d	759dc23452c8a3d3e64e534638e17b7f	fail	7	12/4/2020 2:22:23 PM
00576ef3617e6bbbabfa0b1090e4122d	358b9ab5c1c0f08bc7def310411eb90f	start	7	12/4/2020 2:23:07 PM
00576ef3617e6bbbabfa0b1090e4122d	358b9ab5c1c0f08bc7def310411eb90f	start	7	12/4/2020 2:24:07 PM
00576ef3617e6bbbabfa0b1090e4122d	bd6406065374e1377f1c4270a9c66f1b	start	7	12/4/2020 2:24:39 PM
00576ef3617e6bbbabfa0b1090e4122d	bd6406065374e1377f1c4270a9c66f1b	complete	7	12/4/2020 2:28:52 PM
00576ef3617e6bbbabfa0b1090e4122d	bd6406065374e1377f1c4270a9c66f1b	start	8	12/4/2020 2:29:33 PM

Fig: 6 Entries corresponding to single player

5. Data Preparation:

The dataset from Kwalee looks like the one in **Fig: 6** but extended for many players for a period of one month. To determine the likelihood of failure in a level of the game we need all the levels of the game as an explanatory variable and the predictable variable (start/failure/completion) should be numeric instead of string. We can consider this a multiclass classification problem to predict the class associated with different status i.e. “Start”, “Fail” & “Complete” but starting a level is not much important for our problem statement and so we can ignore that class and treat this as a binary classification problem which fits a Logistic Regression with Logit Link function.

Feature Extraction:

The given data does have only five variables and one of them is our dependant variable, so the more the features we can extract the better the statistical model will be. From the remaining four independent variable, Player_id is not important for our model as it’s only a repeated string variable and so we will ignore that variable. Instead of player_id we can create a cluster to find players who are similar based on the time they play the game. We can also create new variables from the Event_date like time of play in a day, weekday when the player played the game, week number in a year if we have data for more than a year or two. We can also create a new variable explaining the number of times the player failed before reaching a particular level or even the number of sessions taken in every level before failing.

Dummy Variable Encoding:

We use dummy variable encoding instead of One-Hot encoding because dummy variable encoding allows us to prefix the column name for better readability. Since we are dealing with categorical variables the interpretation for statistical significance for other categorical variables will be compared with a reference variable. In case of “level_number” factor, we will keep “Level_number_3” as the reference variable because it has majority of data in the dataset as observed in **Fig: 3** and **Fig: 5.** We did not choose “level_number_1“ or “level_number_2“ as

reference level because we assume those two levels are beginning levels and game developers will not introduce much difficulty in the initial levels thereby encouraging players to engage with the game and gain interest in the game.

Smote-NC: Synthetic Minority Over-sampling Technique for Nominal and Continuous:

SMOTE-NC is a technique used to increase the number of training samples in the training dataset which has both categorical and numerical data, when there is an imbalance in the predictor class. In our case the number of failures is too low and so the statistical model will not be able to produce better results and the model is over fitting towards the status complete as it's the majority portion in the data compared to the failure status. Leaving out the start status in the dataset and plotting the count of Failure status and completion status produces Fig: 7 where we can clearly see the imbalance in the failure status which is marked as 1 in the x-axis.

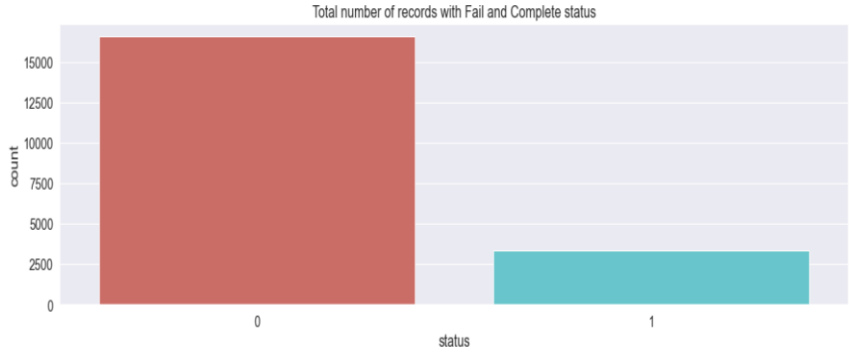


Fig: 7 Class imbalance between status Fail and complete

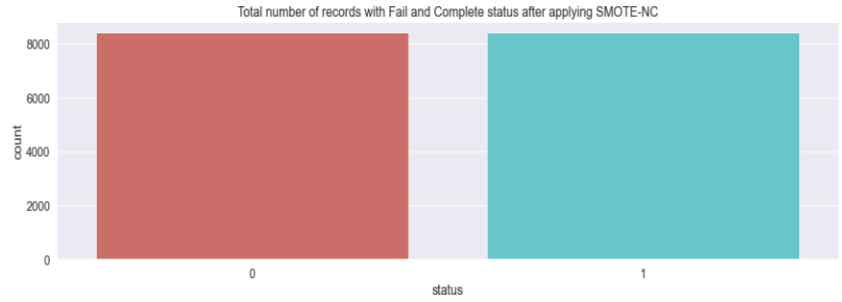


Fig: 8 Status Fail (1) Complete (0) balanced using SMOTE-NC

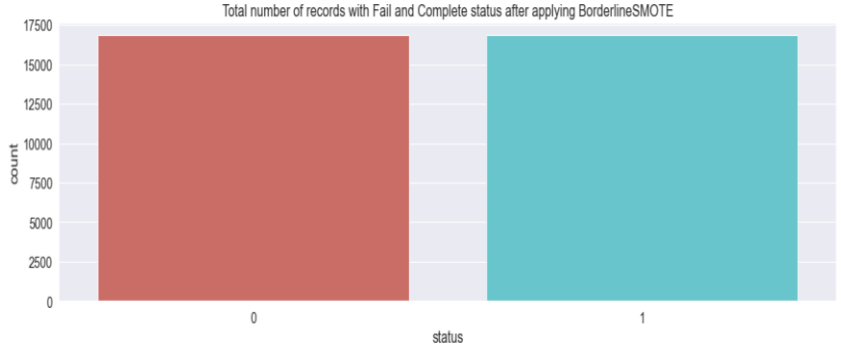


Fig: 9 Status Fail (1) Complete (0) balanced using BorderlineSMOTE

Borderline SMOTE, an adaptive smote technique which avoids over fitting in its process. Towards the end we will compare SMOTE-NC with Borderline SMOTE to choose the best model. From **Fig: 8 & Fig: 9** it's also visible that the number of records imputed through both SMOTE techniques varies.

6. Statistical Modeling

We start our modeling by choosing a **Generalized Linear Model (GLM)** with a **LOGIT Link function** to predict the change of Fail/Completion of a game by several players. For the first model we use the "level_numbers" (from level 1 till level 49) as the only explanatory variable. We stick to Generalized Linear model from the statsmodel package as Sklearn package of Logistic Regression does not support P-values and confidence intervals to validate our model and so we stick with statsmodel package for modeling in the beginning and use Sklearn package for evaluating model accuracy.

Train-Test-split:

The dataset was split into 70% for training and 30% as test set to evaluate our model results. Since we have an imbalanced class on the predictor variable we did a stratified train and test split to have a balance in the given data between the divided sets.

Model 1: Model with all "47 level_numbers" as categorical explanatory variable and status (fail/complete) as predictor. Results from Fig: 10 indicate that the model did not converge even after 5000 iterations. We only used 47 variables as level 0 & 49 did not had a fail or complete status and level 3 is the reference level and so it's excluded from modeling. Choice of choosing level 3 as the baseline is because Level 3 has more observations than other Levels.

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	10261
Model:	GLM	**Df Residuals:**	10261
Model Family:	Binomial	**Df Model:**	-1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Wed, 06 Jan 2021	**Deviance:**	nan
Time:	16:20:15	**Pearson chi2:**	1.19e+19
No. Iterations:	5000		
Covariance Type:	nonrobust		

Fig-10 Model-1 with stratified train test split and 47 explanatory variables

Model 2: After repeated trials we ended with a model that converged which had level-1 till level-29 as the explanatory variable with level-3 as baseline. Inclusion of levels beyond 29 to a model did not converge as a result we decided to stick with level 29 for Model 2. As seen in **Fig: 11,** the model converged in 8 iterations as compared to Model-1 which did not converge even until 5000 iterations which explains that the levels 30 till 49 has very low number of observations compared to others levels, which is also visible in **Fig: 5**.

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	10261
Model:	GLM	**Df Residuals:**	10232
Model Family:	Binomial	**Df Model:**	28
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4226.1
Date:	Wed, 06 Jan 2021	**Deviance:**	8452.2
Time:	14:55:09	**Pearson chi2:**	1.03e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0664	0.253	-4.215	0.000	-1.562	-0.570
level_number_1	-4.2061	0.419	-10.035	0.000	-5.028	-3.385
level_number_2	-2.2261	0.288	-7.733	0.000	-2.790	-1.662
level_number_4	-0.2347	0.263	-0.893	0.372	-0.750	0.280
level_number_5	0.1507	0.262	0.575	0.565	-0.363	0.664
level_number_6	0.2000	0.265	0.755	0.450	-0.319	0.719
level_number_7	-0.0567	0.270	-0.210	0.834	-0.586	0.473
level_number_8	0.2181	0.271	0.805	0.421	-0.313	0.749
level_number_9	-0.1542	0.280	-0.551	0.582	-0.703	0.395
level_number_10	-0.2290	0.287	-0.797	0.425	-0.792	0.334
level_number_11	-0.0194	0.285	-0.068	0.946	-0.578	0.539
level_number_12	0.0819	0.290	0.282	0.778	-0.487	0.651
level_number_13	0.1860	0.291	0.639	0.523	-0.385	0.757
level_number_14	0.1630	0.294	0.554	0.580	-0.414	0.740
level_number_15	-0.2182	0.324	-0.673	0.501	-0.854	0.418
level_number_16	-0.0523	0.323	-0.162	0.871	-0.685	0.580
level_number_17	0.1440	0.326	0.442	0.659	-0.495	0.783
level_number_18	-0.2686	0.357	-0.753	0.451	-0.968	0.430
level_number_19	0.1570	0.335	0.468	0.640	-0.500	0.814
level_number_20	-0.4535	0.389	-1.167	0.243	-1.215	0.308
level_number_21	-0.4918	0.406	-1.211	0.226	-1.288	0.304
level_number_22	0.2496	0.374	0.667	0.505	-0.483	0.983
level_number_23	0.6609	0.365	1.809	0.070	-0.055	1.377
level_number_24	-1.6745	0.772	-2.169	0.030	-3.188	-0.161
level_number_25	-0.8506	0.542	-1.570	0.116	-1.912	0.211
level_number_26	0.1108	0.450	0.246	0.805	-0.771	0.993
level_number_27	-0.2774	0.524	-0.530	0.596	-1.304	0.749
level_number_28	-0.7254	0.673	-1.078	0.281	-2.044	0.594
level_number_29	0.3042	0.523	0.582	0.561	-0.721	1.329

Fig-11 Model-2 with stratified train test split and 47 explanatory variables

The model in Fig:11 explains the significance of each level in the game that corresponds to the likelihood of failure. Since it's a GLM model we have the Constant (Const) term as the first variable. The **Log-Likelihood of a null model with only constant term is -13599.815** where as the 2nd converged model **i.e. Model-2** has a Log-likelihood value of -4226.1**.** As **GLM** works on the principle of maximizing Log-Likelihood estimation, comparing the null model with Model-2 shows that the addition of 28 levels has better parameter estimates than a null model and it cannot be rejected.

Further, the constant term is very much significant at 95% level with a Z value of -4.215 which is defined by the below formula.

$$Z = \frac{Coef}{Std\ Err}$$

Significant values for Z at different Confidence Interval are listed in the below table,

Confidence Interval	z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

In Model-2, the constant term is highly significant indicating that predicting the probability of failure depends on more explanatory variable than just the Level numbers. Compared to Level-3, we observe that estimated coefficient for Level-1 and Level-2 is significantly different from zero at 99.9%. The negative sign indicates the fact that Level-1 & Level-2 has less number of failures that Level-3. The results matches with our expectation, that the initial levels in the game are not that harder than higher levels and so level-1 and Level-2 has lesser chance of failure than Level-3, still players who are new to the game fail at Level-1 and Level-2 which is evident from the significant Z value in Model-2.

The estimated co-efficient of all other Levels are having a p-value higher than 0.05 indicating that they are close to Zero. Looking at the Confidence Interval of the estimated parameters we can see that the value ZERO lies between the upper and Lower bound which is another indication that majority of the Levels estimated Coefficient values are very much close to zero. **Interestingly, the estimated coefficient for Level 24 is significant at 95% level and its Confidence Interval does not have ZERO between the upper and lower range. Negative sign of the coefficient for Level 24 indicates that it's has less chances of failure compared to Level-3, which is our reference variable. With respect to Level 23, its coefficient is significantly different from ZERO at 90% level and the positive sign indicates that a chance of failure in Level 23 is higher when compared to Level 3.**

Model-3: Adding the variable hour of play to Model-2 does not produce significant different in the estimated parameters. Model-3 converged again in 8 iterations and the Log-Likelihood increased by small number with the addition of "Hour" variable. The Hour variable itself is significant at 90% level and the positive sign for coefficient indicate that the failure at different levels increases as more players play during the day than the early hours of the day which is evident from **Fig-4.** Although there is a minor change in the estimates of different levels compared to Model-2, their significance remains the same as in the previous model.

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	10261
Model:	GLM	**Df Residuals:**	10231
Model Family:	Binomial	**Df Model:**	29
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4224.4
Date:	Wed, 06 Jan 2021	**Deviance:**	8448.8
Time:	14:55:09	**Pearson chi2:**	1.03e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2030	0.264	-4.555	0.000	-1.721	-0.685
hours	0.0088	0.005	1.813	0.070	-0.001	0.018
level_number_1	-4.2023	0.419	-10.025	0.000	-5.024	-3.381
level_number_2	-2.2213	0.288	-7.715	0.000	-2.786	-1.657
level_number_4	-0.2262	0.263	-0.861	0.389	-0.741	0.289
level_number_5	0.1607	0.262	0.613	0.540	-0.353	0.675
level_number_6	0.2095	0.265	0.791	0.429	-0.310	0.729
level_number_7	-0.0441	0.270	-0.163	0.870	-0.574	0.486
level_number_8	0.2299	0.271	0.848	0.396	-0.301	0.761
level_number_9	-0.1403	0.280	-0.501	0.617	-0.689	0.409
level_number_10	-0.2152	0.287	-0.749	0.454	-0.778	0.348
level_number_11	-0.0069	0.285	-0.024	0.981	-0.566	0.552
level_number_12	0.0954	0.290	0.329	0.742	-0.474	0.664
level_number_13	0.1931	0.291	0.663	0.507	-0.378	0.764
level_number_14	0.1690	0.294	0.574	0.566	-0.408	0.746
level_number_15	-0.2095	0.324	-0.646	0.518	-0.845	0.426
level_number_16	-0.0457	0.323	-0.142	0.887	-0.679	0.587
level_number_17	0.1472	0.326	0.451	0.652	-0.492	0.787
level_number_18	-0.2594	0.357	-0.727	0.467	-0.958	0.440
level_number_19	0.1700	0.335	0.507	0.612	-0.487	0.827
level_number_20	-0.4441	0.389	-1.142	0.253	-1.206	0.318
level_number_21	-0.4778	0.406	-1.176	0.239	-1.274	0.318
level_number_22	0.2647	0.374	0.707	0.479	-0.469	0.998
level_number_23	0.6661	0.365	1.823	0.068	-0.050	1.382
level_number_24	-1.6581	0.772	-2.147	0.032	-3.172	-0.144
level_number_25	-0.8488	0.542	-1.567	0.117	-1.911	0.213
level_number_26	0.1185	0.450	0.263	0.792	-0.764	1.001
level_number_27	-0.2611	0.524	-0.499	0.618	-1.288	0.765
level_number_28	-0.7347	0.673	-1.091	0.275	-2.054	0.585
level_number_29	0.3036	0.523	0.580	0.562	-0.722	1.329

Fig-12 Model-3 with Hour as an additional explanatory variable to Model-2

Model-4: Experiment with hour of play variable as a categorical one showed some difference in the significance on the constant term but it's not differing much when compared to Mode-3 as the constant term is still significant at 99.9% level. We applied K-means clustering technique to the Hours variable and found 3 clusters as an optimal one for Hour variable. We can see the results of **K-means clustering in Fig: 14** shows some pattern with the Hours of play when compared to **Fig: 4,** there are three distinct time periods where the players play the game. Starting from midnight to early morning the playing time is less and then until 3pm it's increasing gradually and from evening 4pm onwards there is a peak until midnight. Further, the Log-Likelihood value is same as in Model-3 and so we decide to stick to Model-3 and ignore Model-4.

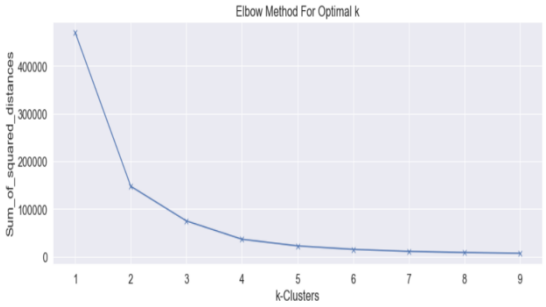


Fig: 14 K-means clustering to find optimal cluster

Fig: 15 clustering result of 3-mean

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	10261
Model:	GLM	**Df Residuals:**	10230
Model Family:	Binomial	**Df Model:**	30
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4224.2
Date:	Wed, 06 Jan 2021	**Deviance:**	8448.4
Time:	14:55:11	**Pearson chi2:**	1.03e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0170	0.254	-3.999	0.000	-1.515	-0.519
level_number_1	-4.2080	0.419	-10.038	0.000	-5.030	-3.386
level_number_2	-2.2285	0.288	-7.739	0.000	-2.793	-1.664
level_number_4	-0.2332	0.263	-0.887	0.375	-0.748	0.282
level_number_5	0.1540	0.262	0.587	0.557	-0.360	0.668
level_number_6	0.2027	0.265	0.765	0.444	-0.317	0.722
level_number_7	-0.0507	0.270	-0.187	0.851	-0.580	0.479
level_number_8	0.2254	0.271	0.831	0.406	-0.306	0.757
level_number_9	-0.1465	0.280	-0.523	0.601	-0.696	0.403
level_number_10	-0.2211	0.287	-0.770	0.441	-0.784	0.342
level_number_11	-0.0151	0.285	-0.053	0.958	-0.574	0.543
level_number_12	0.0895	0.290	0.308	0.758	-0.479	0.659
level_number_13	0.1877	0.291	0.644	0.519	-0.383	0.759
level_number_14	0.1634	0.294	0.555	0.579	-0.414	0.741
level_number_15	-0.2101	0.324	-0.647	0.517	-0.846	0.426
level_number_16	-0.0441	0.323	-0.137	0.891	-0.677	0.589
level_number_17	0.1424	0.326	0.437	0.662	-0.497	0.782
level_number_18	-0.2630	0.357	-0.737	0.461	-0.962	0.436
level_number_19	0.1640	0.335	0.489	0.625	-0.493	0.821
level_number_20	-0.4519	0.389	-1.162	0.245	-1.214	0.310
level_number_21	-0.4800	0.406	-1.182	0.237	-1.276	0.316
level_number_22	0.2545	0.374	0.680	0.496	-0.479	0.988
level_number_23	0.6565	0.365	1.796	0.072	-0.060	1.373
level_number_24	-1.6651	0.772	-2.156	0.031	-3.179	-0.151
level_number_25	-0.8614	0.542	-1.590	0.112	-1.923	0.201
level_number_26	0.1033	0.450	0.229	0.819	-0.779	0.986
level_number_27	-0.2671	0.524	-0.510	0.610	-1.294	0.760
level_number_28	-0.7418	0.673	-1.102	0.271	-2.061	0.578
level_number_29	0.2909	0.523	0.556	0.578	-0.735	1.317
Cluster_1	-0.0940	0.091	-1.035	0.301	-0.272	0.084
Cluster_2	-0.1068	0.058	-1.856	0.064	-0.220	0.006

Fig: 16 Model-4 with Hours as categorical variable

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	10261
Model:	GLM	**Df Residuals:**	10261
Model Family:	Binomial	**Df Model:**	-1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4224.4
Date:	Wed, 06 Jan 2021	**Deviance:**	8448.8
Time:	14:55:12	**Pearson chi2:**	1.03e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4718	0.087	-16.847	0.000	-1.643	-1.301
hours	0.0088	0.005	1.813	0.070	-0.001	0.018
level_number_1	-3.9334	0.327	-12.015	0.000	-4.575	-3.292
level_number_2	-1.9525	0.143	-13.625	0.000	-2.233	-1.672
level_number_4	0.0426	0.087	0.489	0.625	-0.128	0.213
level_number_5	0.4295	0.085	5.031	0.000	0.262	0.597
level_number_6	0.4784	0.093	5.134	0.000	0.296	0.661
level_number_7	0.2247	0.106	2.117	0.034	0.017	0.433
level_number_8	0.4987	0.108	4.613	0.000	0.287	0.711
level_number_9	0.1286	0.128	1.005	0.315	-0.122	0.379
level_number_10	0.0536	0.142	0.378	0.706	-0.225	0.332
level_number_15	0.0593	0.203	0.292	0.771	-0.339	0.458
level_number_16	0.2231	0.201	1.109	0.267	-0.171	0.617
level_number_17	0.4160	0.206	2.019	0.044	0.012	0.820
level_number_18	0.0095	0.249	0.038	0.970	-0.478	0.497
level_number_19	0.4388	0.219	2.001	0.045	0.009	0.869
level_number_20	-0.1753	0.290	-0.604	0.546	-0.744	0.393
level_number_21	-0.2090	0.312	-0.671	0.502	-0.820	0.402
level_number_22	0.5336	0.272	1.964	0.049	0.001	1.066
level_number_23	0.9349	0.260	3.591	0.000	0.425	1.445
level_number_24	-1.3893	0.707	-1.965	0.049	-2.775	-0.004
level_number_25	-0.5800	0.466	-1.245	0.213	-1.493	0.333
level_number_26	0.3873	0.364	1.065	0.287	-0.325	1.100
level_number_27	0.0077	0.446	0.017	0.986	-0.867	0.882
level_number_28	-0.4658	0.605	-0.770	0.441	-1.652	0.720
level_number_29	0.5724	0.446	1.284	0.199	-0.301	1.446
level_number_30	0.2688	0.250	1.073	0.283	-0.222	0.760

Fig: 17 Model-5 includes level 30-49 as aggregated as Level 30

Model 5: Instead of ignoring Level 30 till Level 48 we decided to aggregate those levels and consider them as level 30. With this addition of Level 30, the significance of Constant term increased with the same sign for the coefficient, the estimated coefficients of Level 1 & Level 2 does not exhibit any major changes to its values or sign. However the coefficient estimates of Level 23 is now significant at 99.9% and Level 22 & Level 24 which was not significant at 95% in Model 2 is now significant at 95% level with the addition of Level 30, which is an aggregate of Level 30 till 48. Additionally, estimated coefficients of Level 17 and Level 19 are now significant at 95% level which was not significant without the Level 30 variable. Aslo, Level 5, Level 6, Level 7 & Level 8 have estimated coefficients that are significantly different from zero.

Model 5 estimated Level 1, Level 2, Level 5, Level 6, Level 7, Level 8, Level 12, Level 13, Level 14, Level 17, Level 19, Level 22, Level 23 and Level 24 with coefficients that are significantly different from Zero and the positive coefficients says that Level 23 has highest chance of failure followed by Level 22, Level 8, Level 6, Level 13, Level 19, Level 14, Level 5, Level 17, Level 12 and then Level 7 when they are compared to Level 3. While Level 1, Level 2 and Level 24 in their respective orders have a lower chance of failure when compared to Level 3. The estimated coefficients of the aggregated levels (i.e. Level 30) are not significant.

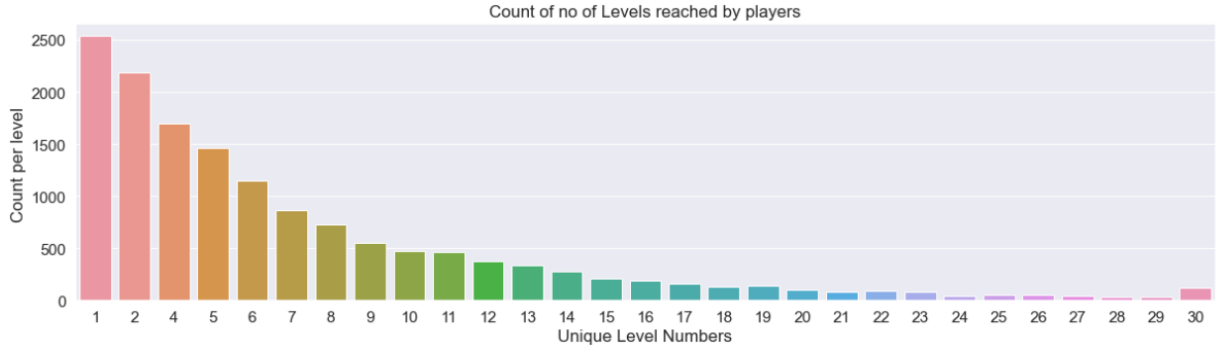


Fig: 18 Aggregate of Levels 30-48 is named as Level 30 which has some visible count for estimation

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	10261
Model:	GLM	**Df Residuals:**	10226
Model Family:	Binomial	**Df Model:**	34
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4221.8
Date:	Wed, 06 Jan 2021	**Deviance:**	8443.7
Time:	14:55:12	**Pearson chi2:**	1.03e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2687	0.076	-16.780	0.000	-1.417	-1.120
hours	0.0089	0.005	1.827	0.068	-0.001	0.018
level_number_1	-3.9316	0.327	-12.009	0.000	-4.573	-3.290
level_number_2	-1.9506	0.143	-13.598	0.000	-2.232	-1.669
level_number_4	0.0483	0.087	0.553	0.580	-0.123	0.220
level_number_5	0.4342	0.086	5.069	0.000	0.266	0.602
level_number_6	0.4848	0.093	5.187	0.000	0.302	0.668
level_number_7	0.2264	0.106	2.128	0.033	0.018	0.435
level_number_8	0.5014	0.108	4.627	0.000	0.289	0.714
level_number_9	0.1325	0.128	1.034	0.301	-0.119	0.383
level_number_10	0.0561	0.142	0.395	0.693	-0.223	0.335
level_number_11	0.2642	0.138	1.918	0.055	-0.006	0.534
level_number_12	0.3667	0.148	2.482	0.013	0.077	0.656
level_number_13	0.4593	0.150	3.069	0.002	0.166	0.753
level_number_14	0.4317	0.155	2.777	0.005	0.127	0.736
level_number_15	0.0559	0.204	0.274	0.784	-0.343	0.455
level_number_16	0.2216	0.201	1.101	0.271	-0.173	0.616
level_number_17	0.4043	0.207	1.958	0.050	-0.000	0.809
level_number_18	0.0032	0.249	0.013	0.990	-0.485	0.491
level_number_19	0.4398	0.219	2.005	0.045	0.010	0.870
level_number_20	-0.1779	0.290	-0.613	0.540	-0.747	0.391
level_number_21	-0.2006	0.312	-0.643	0.520	-0.812	0.410
level_number_22	0.5527	0.272	2.033	0.042	0.020	1.086
level_number_23	0.9606	0.261	3.686	0.000	0.450	1.471
level_number_24	-1.3799	0.707	-1.952	0.051	-2.766	0.006
level_number_25	-0.5586	0.466	-1.198	0.231	-1.472	0.355
level_number_26	0.4024	0.364	1.106	0.269	-0.311	1.116
level_number_27	0.0301	0.446	0.067	0.946	-0.845	0.905
level_number_28	-0.4384	0.605	-0.724	0.469	-1.625	0.748
level_number_29	0.6017	0.446	1.349	0.177	-0.273	1.476
level_number_30	0.2908	0.251	1.159	0.246	-0.201	0.782
day_of_week_num_0	-0.1818	0.058	-3.108	0.002	-0.296	-0.067
day_of_week_num_1	-0.2313	0.061	-3.783	0.000	-0.351	-0.111
day_of_week_num_2	-0.2471	0.065	-3.799	0.000	-0.375	-0.120
day_of_week_num_3	-0.3165	0.066	-4.804	0.000	-0.446	-0.187
day_of_week_num_4	-0.1142	0.062	-1.830	0.067	-0.237	0.008
day_of_week_num_5	-0.1778	0.059	-3.013	0.003	-0.293	-0.062

Fig: 19 Model 6 with Day_of_week_number as a categorical variable

Model-6: Day of the Week has been added to Model-5 and **Fig: 19** show the result of the statistical estimates where there is no major changes to the estimates of different levels when compared to Model-5. But the Day_of_the_week categorical variable is all significant indicating that a fail or completion chances does not have any affinity with a particular day in the week. With Sunday being our reference variable, the negative sign for coefficients states that when compared to Sunday, day_of_week_num_5 (Saturday) has next higher change of Failure, followed by day_of_week_num_1 (Monday), day_of_week_num_2 (Tuesday) and day_of_week_num_3 (Wednesday). Day_of_week_num_4 (Thursday) is not significant at 95% level however it's significant at 90% level.

The Log-Likelihood value of Model-6 is -4221.8 which is slightly higher than Model-5 which was -4224.4 indicating the positive effect of adding a new variable in the statistical estimates.

Starting from Model-1 till Model-6, we used a dataset which had an imbalance in the Predictor variable. As seen in the **Fig: 20**, the train set and test set both had an imbalance in the Status (fail/Complete) variable. However because of stratified train test split, we managed to get a balanced proportion of Status =1 in train and test set yet the imbalance is clearly visible in **fig: 20**.

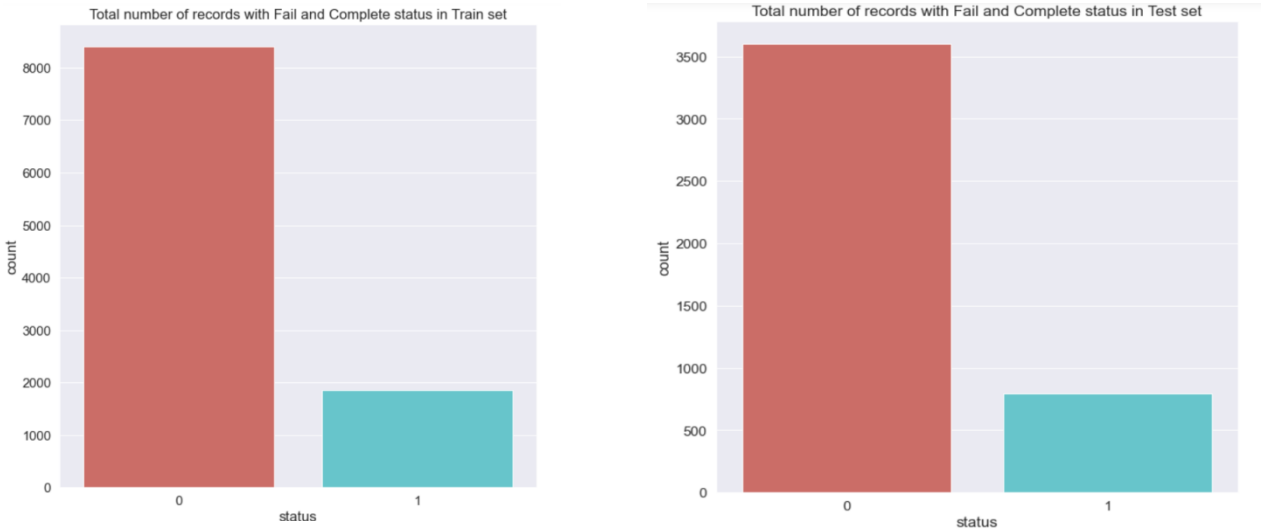


Fig: 20 Stratified Train Test split of given dataset used for Model 6

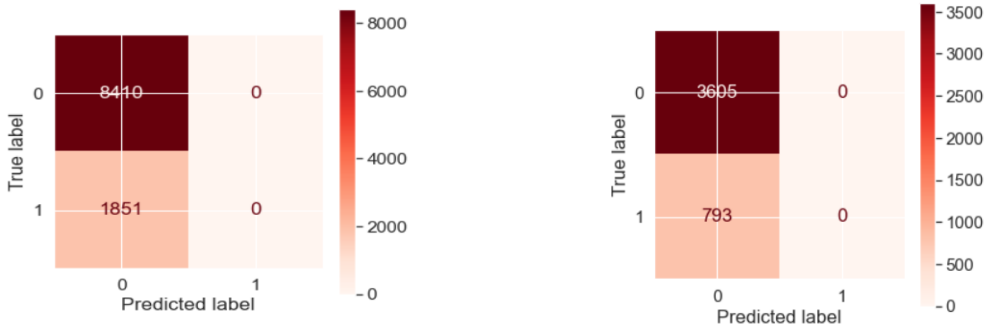


Fig: 21 Confusion Matrix for Model 6 on Train set and Test set

Testing the Model-6 on Test set revealed that the model only predicts Completed status which is the majority class in the predictor variable. So we cannot rely on Model-6 results though the statistical estimates are interesting.

To overcome the problems with imbalanced dataset used in Model-6, we introduce SMOTE techniques to increase the number of samples to have a balance between the Status - Fail and Complete. In the next section we will experiment SMOTE-NC and Borderline SMOTE techniques to estimate the model parameters. Borderline SMOTE technique is an adaptive model which avoids over fitting caused by increase in the sample size.

Model-7: Using the dataset of Model-6 we applied **SMOTE-NC,** which handles both Categorical and Continuous Variables.

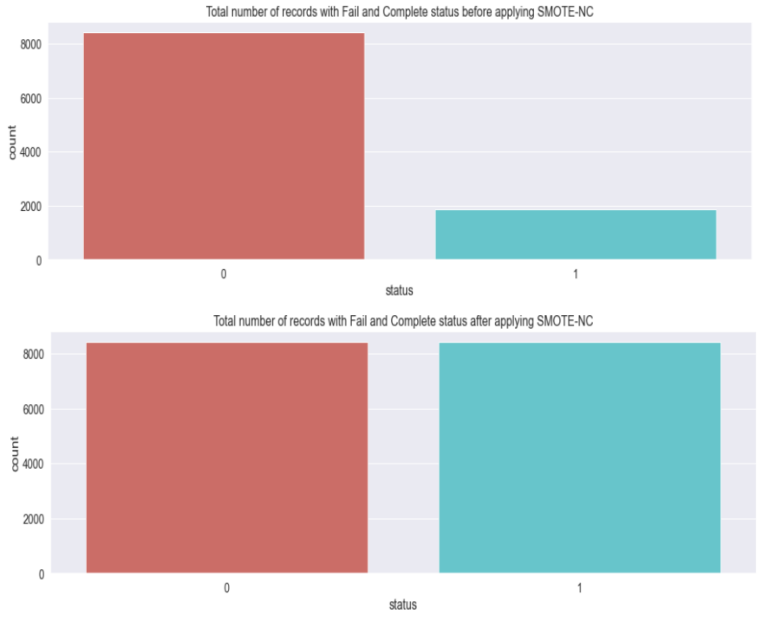


Fig: 22 Split of Status variable before and after applying SMOTE-NC

Results of Model 7 shows that all the Levels (from 1-48 excluding 3) are statistically significant which is great result. However the negative sign on the entire estimated Coefficient indicates that Level 3 has the highest failure chances of all the levels which raise more doubts about the model estimate. The reason could be a possible over fitting because of new samples added. Also the variable day_of_week_num variables have a high standard error making the Model 7 estimates unreliable.

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	16821
Model:	GLM	**Df Residuals:**	16821
Model Family:	Binomial	**Df Model:**	-1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-6893.2
Date:	Wed, 06 Jan 2021	**Deviance:**	13786.
Time:	14:55:18	**Pearson chi2:**	1.71e+04
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	29.9692	1.25e+04	0.002	0.998	-2.46e+04	2.46e+04
hours	0.0206	0.004	5.383	0.000	0.013	0.028
level_number_1	-12.9294	1.031	-12.537	0.000	-14.951	-10.908
level_number_2	-11.5927	1.010	-11.474	0.000	-13.573	-9.613
level_number_4	-8.1639	1.001	-8.154	0.000	-10.126	-6.202
level_number_5	-7.7111	1.001	-7.701	0.000	-9.674	-5.749
level_number_6	-7.8544	1.002	-7.841	0.000	-9.818	-5.891
level_number_7	-8.5729	1.003	-8.549	0.000	-10.538	-6.607
level_number_8	-8.1911	1.003	-8.167	0.000	-10.157	-6.225
level_number_9	-9.0742	1.006	-9.023	0.000	-11.045	-7.103
level_number_10	-8.8407	1.006	-8.790	0.000	-10.812	-6.869
level_number_11	-8.5772	1.006	-8.529	0.000	-10.548	-6.606
level_number_12	-9.0350	1.009	-8.953	0.000	-11.013	-7.057
level_number_13	-8.3255	1.007	-8.271	0.000	-10.299	-6.353
level_number_14	-8.6453	1.008	-8.573	0.000	-10.622	-6.669
level_number_15	-9.0740	1.015	-8.941	0.000	-11.063	-7.085
level_number_16	-8.9318	1.015	-8.801	0.000	-10.921	-6.943
level_number_17	-8.3632	1.014	-8.251	0.000	-10.350	-6.377
level_number_18	-9.2656	1.025	-9.040	0.000	-11.275	-7.257
level_number_19	-9.2585	1.027	-9.015	0.000	-11.271	-7.246
level_number_20	-9.8348	1.046	-9.407	0.000	-11.884	-7.786
level_number_21	-9.6223	1.047	-9.191	0.000	-11.674	-7.570
level_number_22	-8.6393	1.029	-8.394	0.000	-10.657	-6.622
level_number_23	-8.7675	1.039	-8.438	0.000	-10.804	-6.731
level_number_24	-9.8410	1.095	-8.983	0.000	-11.988	-7.694
level_number_25	-9.9754	1.111	-8.977	0.000	-12.153	-7.798
level_number_26	-9.2745	1.066	-8.701	0.000	-11.364	-7.185
level_number_27	-9.4441	1.103	-8.564	0.000	-11.605	-7.283
level_number_28	-10.5324	1.246	-8.454	0.000	-12.974	-8.091
level_number_29	-8.9311	1.101	-8.114	0.000	-11.088	-6.774
level_number_30	-8.2103	1.017	-8.072	0.000	-10.204	-6.217
day_of_week_num_0	-21.9999	1.25e+04	-0.002	0.999	-2.46e+04	2.46e+04
day_of_week_num_1	-22.0732	1.25e+04	-0.002	0.999	-2.46e+04	2.46e+04
day_of_week_num_2	-22.1139	1.25e+04	-0.002	0.999	-2.46e+04	2.46e+04
day_of_week_num_3	-22.3116	1.25e+04	-0.002	0.999	-2.46e+04	2.46e+04
day_of_week_num_4	-21.9714	1.25e+04	-0.002	0.999	-2.46e+04	2.46e+04
day_of_week_num_5	-21.9634	1.25e+04	-0.002	0.999	-2.46e+04	2.46e+04

Fig: 23 Model- 7 with SMOTE-NC

Model 8: Since SMOTE_NC results were note reliable, we moved to Borderline SMOTE which avoids over fitting from the imputed data. From f**ig: 24** we can see that Border Line smote imputed more records than SMOTE_NC.

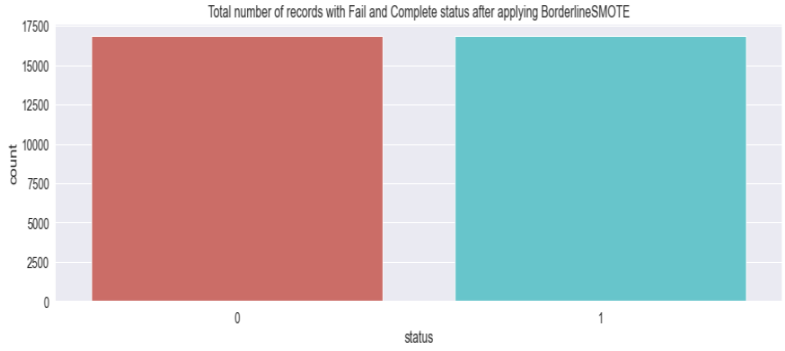


Fig: 24 Borderline SMOTE

Generalized Linear Model Regression Results

Dep. Variable:	status	**No. Observations:**	16085
Model:	GLM	**Df Residuals:**	16085
Model Family:	Binomial	**Df Model:**	-1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-9147.1
Date:	Wed, 06 Jan 2021	**Deviance:**	18294.
Time:	19:27:40	**Pearson chi2:**	1.61e+04
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0164	0.048	-0.340	0.734	-0.111	0.078
hours	0.0026	0.003	0.822	0.411	-0.004	0.009
level_number_1	-3.7626	0.154	-24.371	0.000	-4.065	-3.460
level_number_2	-2.3263	0.089	-26.042	0.000	-2.501	-2.151
level_number_4	0.1074	0.054	1.975	0.048	0.001	0.214
level_number_5	0.6046	0.055	11.044	0.000	0.497	0.712
level_number_6	0.4252	0.062	6.830	0.000	0.303	0.547
level_number_7	0.1331	0.070	1.905	0.057	-0.004	0.270
level_number_8	0.5732	0.072	7.973	0.000	0.432	0.714
level_number_9	0.1456	0.082	1.775	0.076	-0.015	0.306
level_number_10	0.1062	0.090	1.181	0.238	-0.070	0.283
level_number_11	0.5017	0.087	5.789	0.000	0.332	0.672
level_number_12	0.3236	0.099	3.259	0.001	0.129	0.518
level_number_13	0.6657	0.098	6.794	0.000	0.474	0.858
level_number_14	0.4015	0.105	3.806	0.000	0.195	0.608
level_number_15	0.1147	0.129	0.891	0.373	-0.138	0.367
level_number_16	0.3601	0.129	2.794	0.005	0.108	0.613
level_number_17	0.2057	0.145	1.415	0.157	-0.079	0.491
level_number_18	0.0675	0.157	0.430	0.667	-0.240	0.375
level_number_19	0.5455	0.146	3.737	0.000	0.259	0.832
level_number_20	-0.6109	0.204	-2.993	0.003	-1.011	-0.211
level_number_21	0.0735	0.180	0.407	0.684	-0.280	0.427
level_number_22	0.8362	0.179	4.661	0.000	0.485	1.188
level_number_23	0.9734	0.191	5.087	0.000	0.598	1.348
level_number_24	-0.8104	0.313	-2.589	0.010	-1.424	-0.197
level_number_25	-0.3784	0.263	-1.437	0.151	-0.895	0.138
level_number_26	0.4951	0.243	2.042	0.041	0.020	0.970
level_number_27	-0.4988	0.329	-1.514	0.130	-1.144	0.147
level_number_28	-0.1396	0.334	-0.418	0.676	-0.794	0.515
level_number_29	0.5604	0.316	1.776	0.076	-0.058	1.179
level_number_30	0.2907	0.167	1.742	0.082	-0.036	0.618
day_of_week_num_0	0.0014	0.039	0.035	0.972	-0.074	0.077
day_of_week_num_1	0.0127	0.040	0.316	0.752	-0.066	0.091
day_of_week_num_2	-0.0662	0.043	-1.550	0.121	-0.150	0.018
day_of_week_num_3	-0.1689	0.043	-3.895	0.000	-0.254	-0.084
day_of_week_num_4	0.1103	0.041	2.669	0.008	0.029	0.191
day_of_week_num_5	0.0943	0.038	2.450	0.014	0.019	0.170

Fig: 25 Model-8 using Borderline SMOTE

Model-8 estimates looks promising when compared to Model-7 and it converged in 7 iterations.

- Constant term is insignificant, which is a good start. This explains that the explanatory variables used in Model-8 captures all the effects of Failure or Completion status.
- Hour variable is insignificant, which is also understandable when we look into **fig: 4**, a player fail or complete irrespective of the time of the day.
- Coefficient estimates of Level-1 and Level-2 are highly significant. The negative sign indicates that Level-1 and Level-2 has less chance of failure than Level-3.
- Coefficient estimates for Levels- 4, 5, 6, 8, 11, 12, 13, 14, 16, 19, 22, 23 and 26 are statistically significant at 95% level. Their Confidence intervals confirms this significance with the absence of ZERO between the upper and lower bound. Positive value of estimates indicate that compared to Level 3 these thirteen levels has a higher chance of failure for players based on the samples imputed through Borderline SMOTE.
- Compared to Level 3- Level 23 has the highest chance of failure among all the levels with a estimate value 0.9734, Levels 22 falls next with an estimate of 0.8362, followed by Level 13,5,8,19,11,26,6,14,16,12,4 whose estimates are 0.9734, 0.8362, 0.6657, 0.6046, 0.5732, 0.5455, 0.5017, 0.4951, 0.4252, 0.4015, 0.3601, 0.3236 and 0.1074 respectively.
- Level 20 and Level 24 are also significant but their negative sign indicates that chances of failure are lesser than Level 3 and even lower rate of failure than Level-1 and Level-2.
- With respect to day_of_week_num, after applying Borderline SMOTE, Friday and Saturday has higher failure rate than Sunday where as Wednesday has less failure rate than Sunday. The estimated coefficients indicate that the difference between Wednesday, Friday, Saturday and Sunday are not very high.

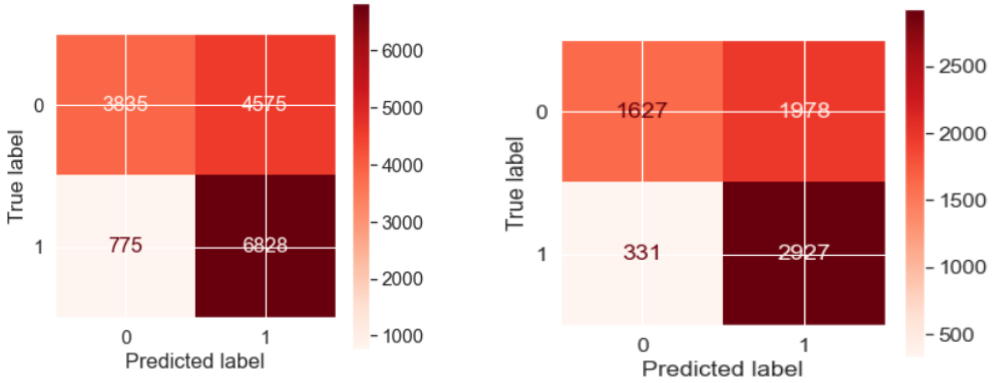
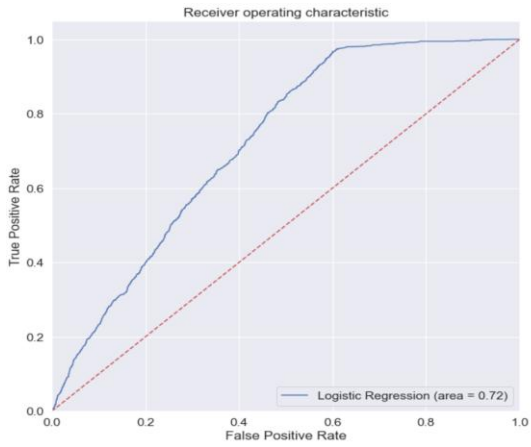


Fig: 26 Confusion Matrix of Training set and Test Set of Model-8

Results of Confusion matrix from Fig: 26 show that Model-8 is capable of predicting both Fail and Complete status. **Fig: 27** show that Model-8 is not randomly predicting the Status of Players playing the game.



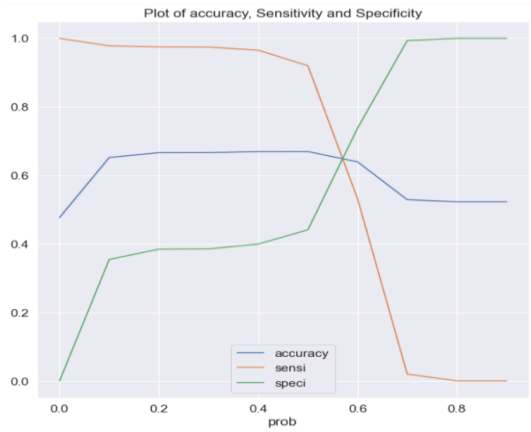


Fig: 27 ROC curve and Accuracy plot of Model-8

7. Conclusion:

Based on the experiments performed, Model 8 seems to emerge out as a clear favorite. Although Model-8 does not have a very high accuracy, the statistical estimate of coefficients indicates that the model estimates are favoring our expectations.

Question: On which level are players most likely to fail?

Answer: Level 23 clearly emerged as the level with highest chance of failure of all followed by Level 22, Level 13, Level 5, Level 8, Level 19, Level 11, Level 26, Level 6, Level 14, Level 16, Level 12 and Level 4 when compared with Level 3 which is our reference class to our comparison. Level 1 and Level 2 has less chance of failure followed by Level 20 and Level 24 compared to Level 3.

Note: Results in **fig: 26** & **fig: 27** can change in every execution of the Jupiter notebook (attached in the location as this file) because of BorderLine SMOTE which imputes new samples to create balance in the given dataset.

The results are not final and there is still scope for improvement to Model-8.

8. Future Works

We started our analysis with only 5 variables and were able to get a reasonable estimate using Borderline SMOTE. It will be good to test the model with more explanatory variables like, difficulty indicators in the game, no of obstacles in each level, whether it's a male or female player, whether the game was played on IOS or Android, Screen resolution of mobile and many other characteristics.

Also with the existing data, still we can extract new features like no of failures in the previous levels, no of sessions the person took to complete previous levels.

Moreover, the existing data has many inconsistencies as visible in **Fig: 6,** where the start time is after fail time, in some cases a player had played the same level twice in different week and completed the game

both times resulting in two entries from same player. Correcting these **inconsistencies** and applying **Cross validating** to our model using a Validation set can also improve the model accuracy.

Thanks Kwalee, it was fun to work in this exercise and thanks again for the opportunity to work on your company's dataset!