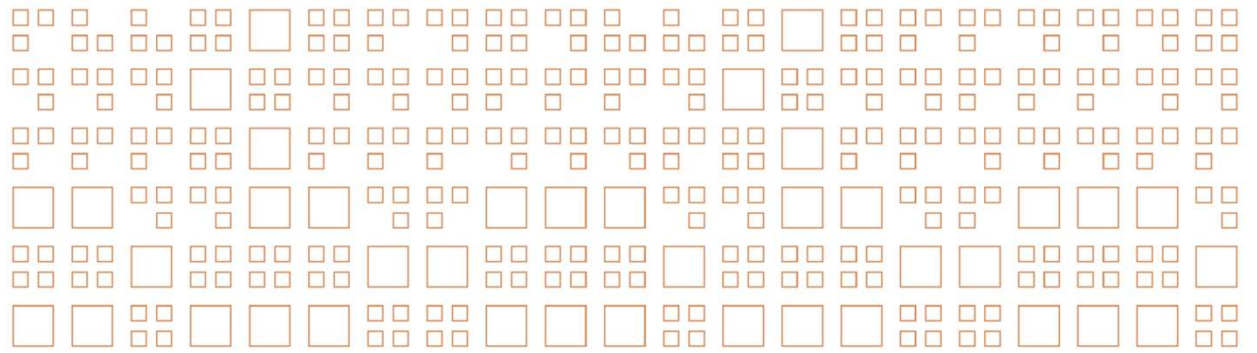


# Predictive Modelling for 2023SP Course Cancellation Analysis

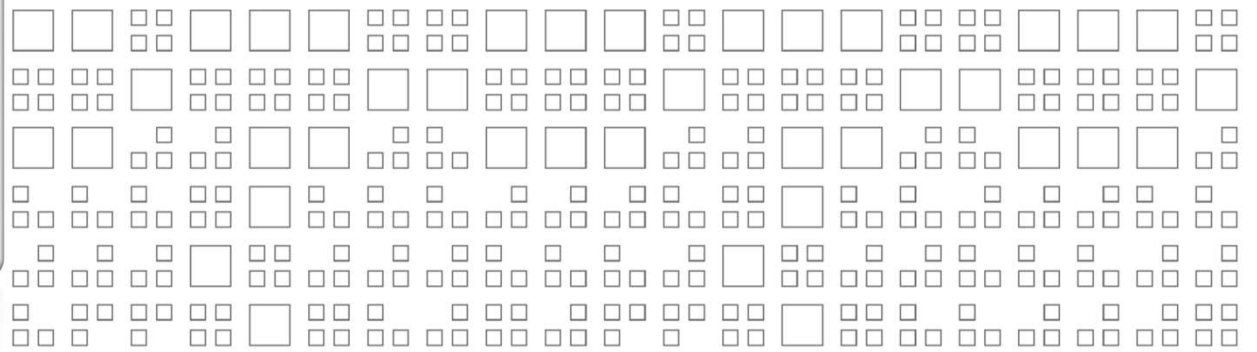
**Daniel Le (Assistant Director)**  
**Dr. Jeremy Monteath-Valdez (Assistant Director)**



PART I:



# INTRODUCTION

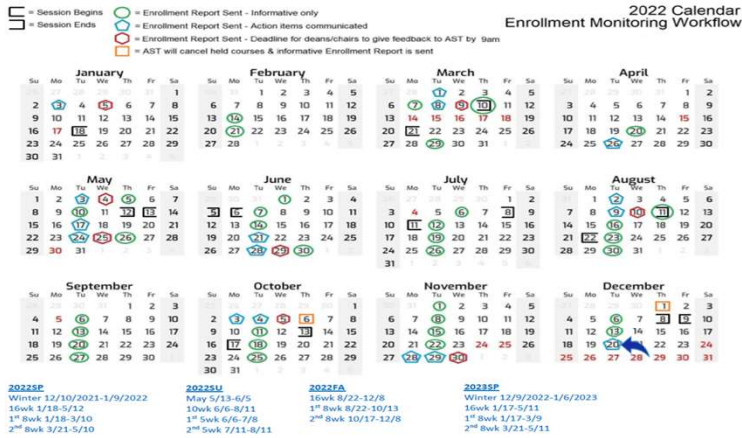




# Brief Introduction

- Academic Scheduling Team (AST) is the scheduling department at Dallas College who is responsible to create Credit sections schedule for the entire college.
- We are working closely with school leadership (Vice Provosts, Deans, Chairs) to monitor the enrollment trend, assign faculty assignment, book adoption, open or cancel sections under the school leadership's requests.
- During the enrollment monitoring period (one month prior to the semester starts), AST is tracking daily enrollment and advise school leadership to cancel low enrolled sections or create new section when the enrollment grows significantly.





**2023 Calendar Enrollment Monitoring Workflow**



Informative Enrollment Report  
Sent - 12/13/22

AST sent the informative enrollment report to school leadership.

Enrollment Monitoring  
Start Date - 12/20/22

AST sent the **action required enrollment report** to school leadership.

Original Cancellation  
Deadline - 01/05/23

**9 AM on Thursday (01/05/23)**, deans/chairs identify courses to be cancelled OR put in watchlist.

Extended Cancellation  
Deadline - 01/09/23

**4 PM on Monday (01/09/23)** to cancel watchlist courses.

**2023SP 16WK & FIRST 8WK CANCELLATION PROCESS TIMELINE**



# Inclusive Criteria

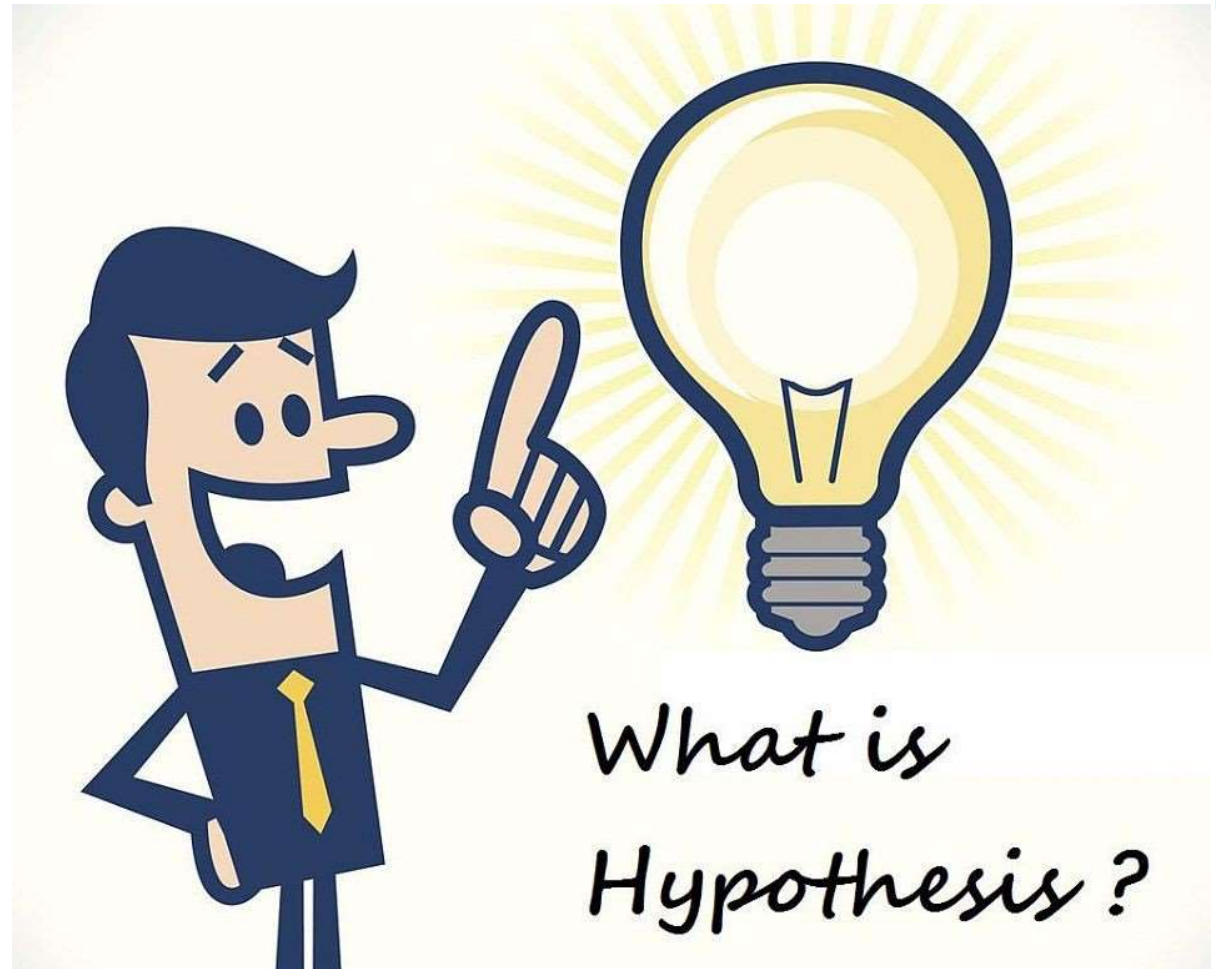
- This cancellation deadline only applies to Credit (CR) sections which start on 01/17/23.
- On the action required enrollment report sent on 12/20/22, we identified 2,021 Active CR sections (325 Dual Credit, 1,696 Gen Pop) which were either empty or under enrolled. This is the dataset of our analysis.





1. Does the extension period have a statistically significant impact on the number of cancelled sections?
2. Will highly-enrolled classes benefit from the extension period?
3. Which other factors are statistically significant in explaining the cancellation rate?

To answer these questions, we will be applying a predictive modelling (machine learning) technique, namely logistic regression.





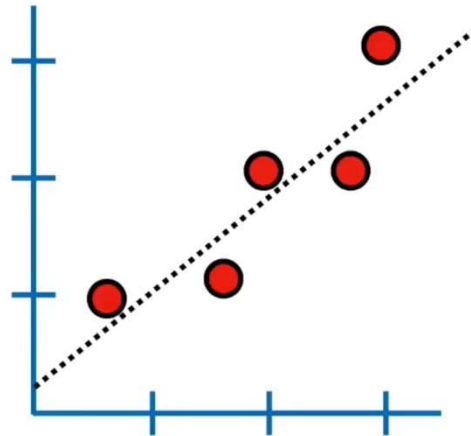
Let's watch a brief video explaining Logistic Regression!

*Regression or Classification?*

*Oh God! What is this LOGISTIC?*

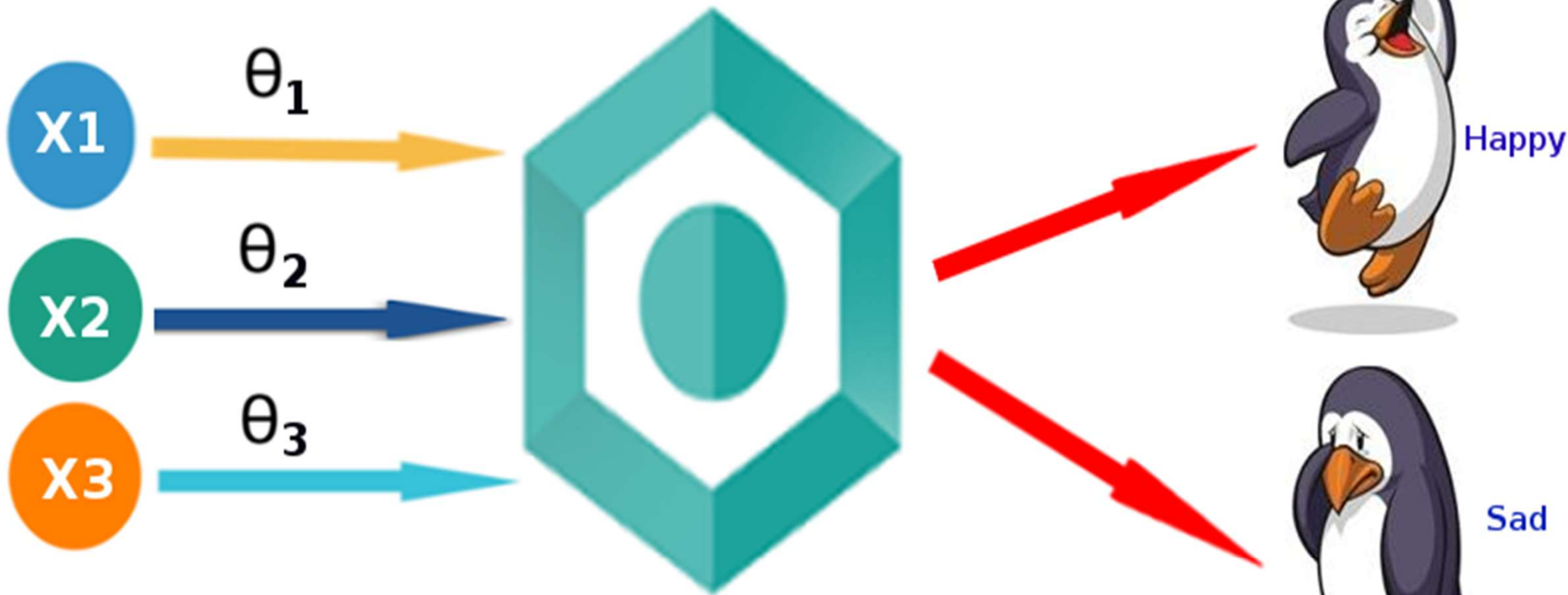


**If you can fit a line...**





# Logistic Regression Model



Inputs:  $X_1, X_2, X_3$  || Weights:  $\theta_1, \theta_2, \theta_3$  || Outputs: Happy or Sad

DL\_JV

[@dataaspirant.com](https://dataaspirant.com)

2/3/23

9



# Original Deadline (01/05/23) Variable Description

Variable Name	Variable Description
ENRL_5_DAYS_PRIOR_TO	The number of students who enrolled during the 5 days period prior to 01/05/23.
12.20.22_DOUBLE_DIGIT	Coded as 0 if enrollment on 12/20/22 for a section is a single digit (less than 10), and 1 for a double digit.
GROUP_ENRL_5_DAYS_PRIOR	Coded as 0 if the total enrollment 5 days prior to 01/05/23 is less than 5, and 1 if the total enrollment is $\geq 5$ .
STRT_DATE_ENRL_GROUP	Enrollment on 12/20/22, coded as 0 if enrollment is 0-5, coded as 1 if enrollment is 6-10, coded as 2 if enrollment is $>10$ .
ENRL_GROUP_ORG	Enrollment on 01/05/23, coded as 0 if enrollment is 0, coded as 1 if enrollment is 1-5, coded as 2 if enrollment is 6-10, coded as 3 if enrollment is $>10$ .
ONE_DAY_BEFORE_DEADLINE	Coded as 0 if enrollment on 01/04/23 for a section is a single digit (less than 10), and 1 for a double digit.
TIME_TO_CANCEL_ORIG *	The number of days from 12/20/22 to the day a section is cancelled. If the section is not cancelled during this period, we assign the value 16. <b>*We create this variable to answer question 1 in page 6.</b>
HIGHLY_ENRL_ORG **	Coded as 1 if the average daily enrollment between 12/20/22 and 01/05/23 is $\geq 0.5$ , and 0 if the average is $< 0.5$ . <b>**We create this variable to answer question 2 in page 6.</b>
WATCHLIST	Coded as 1 if the section is on the watchlist, and 0 if not.
LEC_STRT_TIME_FACTOR	Coded as 0 if the lecture start time is between 8AM and 10AM, coded as 1 if the LEC start time is between 10AM and 12PM, coded as 2 if the LEC start time is between 12PM and 2PM, coded as 3 if the LEC start time is between 2PM and 4PM, coded as 4 if the LEC start time is after 4PM, coded as 5 for online classes.



## Model 1: Original Deadline Response Variable Profile

Response Profile		
Ordered Value	EVENT_ORIG	Total Frequency
1	0	1684
2	1	337

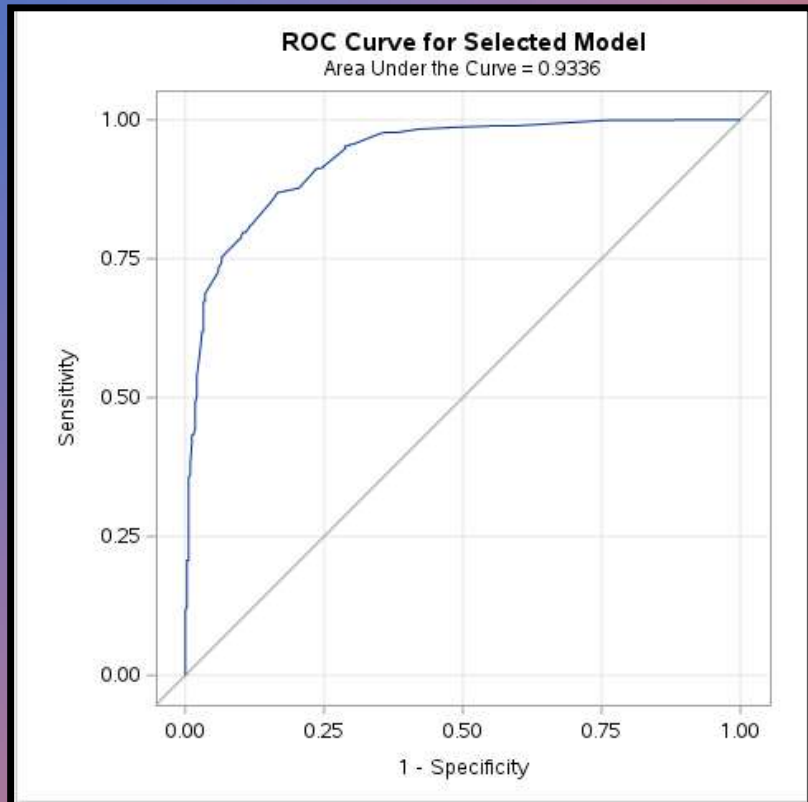
Probability modeled is EVENT\_ORIG='0'.

- In our analysis, EVENT\_ORIG, otherwise known as the response (outcome) variable, is coded as:
  - **0**: The section is NOT cancelled by 01/05/23. This is our event of interest.
  - **1**: The section is cancelled by 01/05/23.

# Model 1 – Original Deadline (01/05/23)

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.6996	0.2403	8.4740	0.0036
ENRL_5_DAYS_PRIOR_TO		1	0.5159	0.0808	40.7468	<.0001
ENRL_GROUP_ORG	1	1	2.3094	0.1937	142.1435	<.0001
ENRL_GROUP_ORG	2	1	4.3614	0.3162	190.2697	<.0001
ENRL_GROUP_ORG	3	1	2.7540	0.9818	7.8685	0.0050
ONE_DAY_BEFORE_DEADL	1	1	3.2814	1.1341	8.3712	0.0038
WATCHLIST	1	1	2.6094	0.5831	20.0288	<.0001

- After applying the backward stepwise selection technique, we obtain the final model that includes these significant factors (p-value < 0.05) at the 95% confidence level.



# Accuracy Level of Model 1

- Area Under the Curve (AUC) is an effective way to summarize the overall diagnostic accuracy of the model. It takes values from 0 to 1, where a value of 0 indicates a perfectly inaccurate test and a value of 1 reflects a perfectly accurate model.
- The final model 1 has the AUC is 0.9336, which implies a **93.36%** chance that the model identifies which sections would be cancelled correctly.
- Let's interpret the impact of the significant factors in explaining the cancellation rate by using the **odds ratio**.

# Odds ratio

- ❖ When  $\theta$  is a probability, the quantity  $\theta/(1-\theta)$  is called odds. The concept of odds has two forms.
- ❖ Suppose  $\theta$  is a probability of “success”.

❖ 1. We define Odds in favor of success =  $\frac{P(\text{success})}{1-P(\text{success})} = \frac{\theta}{1-\theta}$ .

❖ 2. We define Odds against success =  $\frac{1-P(\text{success})}{P(\text{success})} = \frac{1-\theta}{\theta}$ .

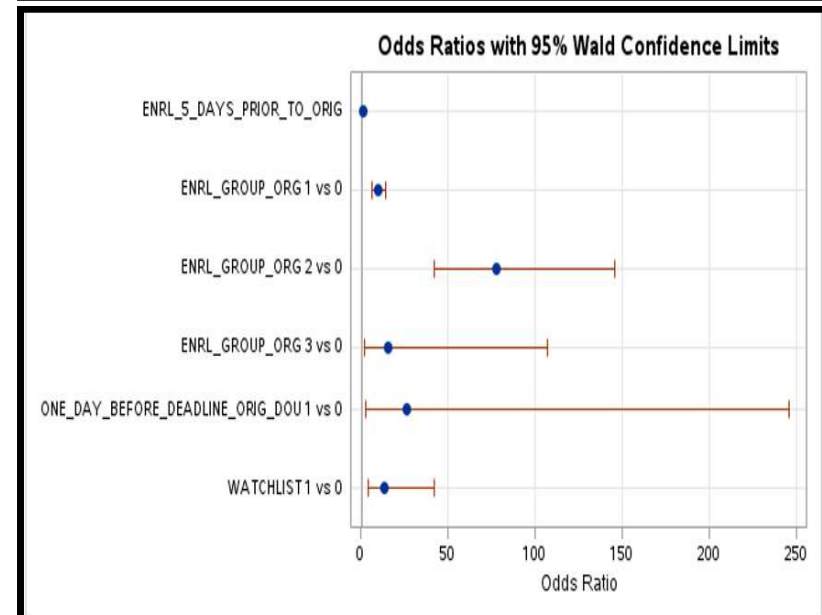
- An **odds ratio (OR)** is a measure of association between a certain property A and a second property B in a population. Specifically, it tells us how the presence or absence of property A influences the presence or absence of property B.
- **An odds ratio of exactly 1** means that exposure to property A does not affect the odds of property B.
- **An odds ratio of more than 1** means that there is a higher odds of property B happening with exposure to property A.
- **An odds ratio is less than 1** is associated with lower odds.



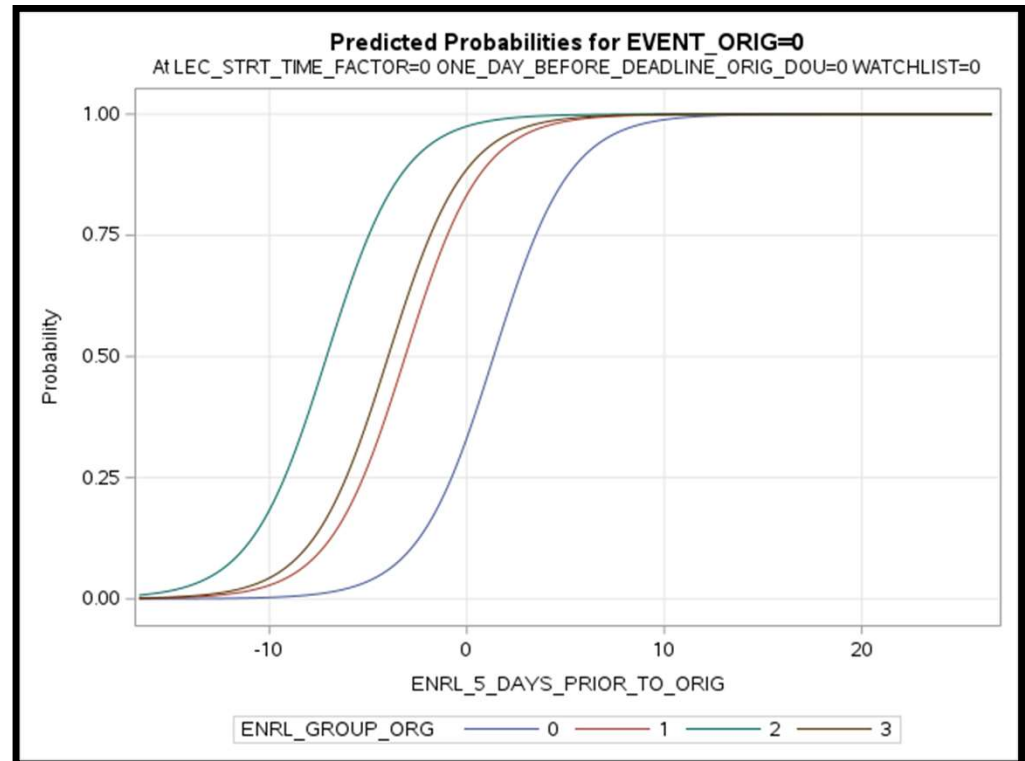
# Model 1 – Final Model's Odds Ratios

- Event of interest is NOT being cancelled by the Original Deadline (01/05/23).
- Assuming other factors stay constant:
  - ENRL\_5\_DAYS\_PRIOR\_TO has Odds Ratio (OR) is **1.675**, which means for **each increment in enrollment during the 5 days period prior to 01/05/23**, the odds of NOT being cancelled of that section is **1.675 times** the odds of other sections.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
ENRL_5_DAYS_PRIOR_TO	1.675	1.430	1.963
ENRL_GROUP_ORG 1 vs 0	10.089	6.888	14.718
ENRL_GROUP_ORG 2 vs 0	78.365	42.168	145.632
ENRL_GROUP_ORG 3 vs 0	15.705	2.293	107.580
ONE_DAY_BEFORE_DEADL 1 vs 0	26.614	2.882	245.754
WATCHLIST 1 vs 0	13.591	4.335	42.615

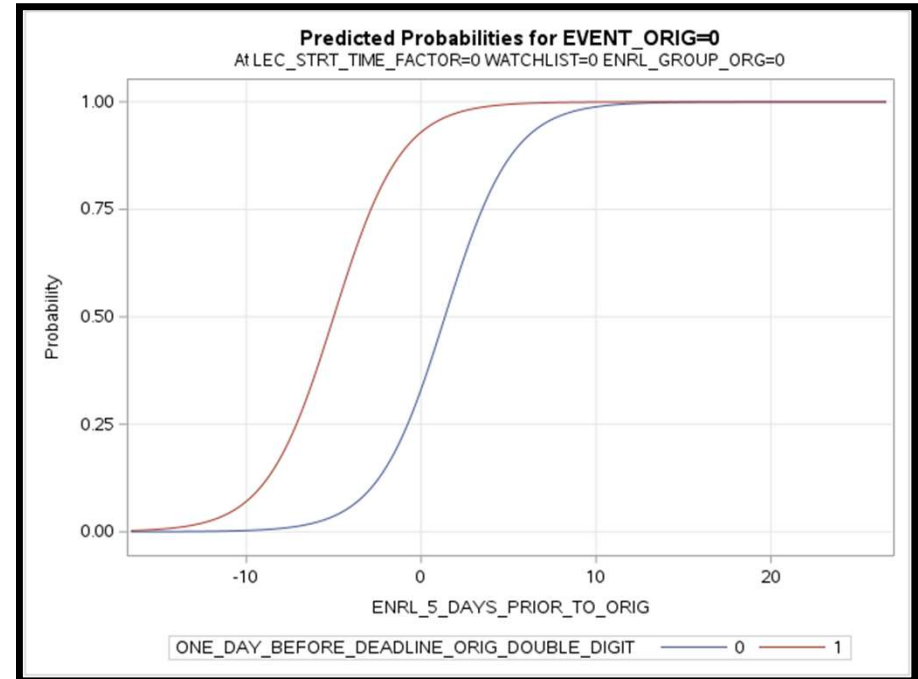


- We plot **ENRL\_5\_DAYS\_PRIOR\_TO\_ORIG** and **ENRL\_GROUP\_ORG** into the same plot for a better visualization of OR.
- For each increment of enrollment during the five days period prior to 01/05/23, **sections with more than 10 enrollment on 01/05/23** (coded as 3) have a higher chance (**15.705** times) of NOT being cancelled than empty sections (coded as 0). The odds of NOT being cancelled for **the 6–10 enrollment group** (coded as 2) is **78.365** times, and for **the 0–5 enrollment group** (coded as 1), the odds are **10.069** times higher than those empty sections.

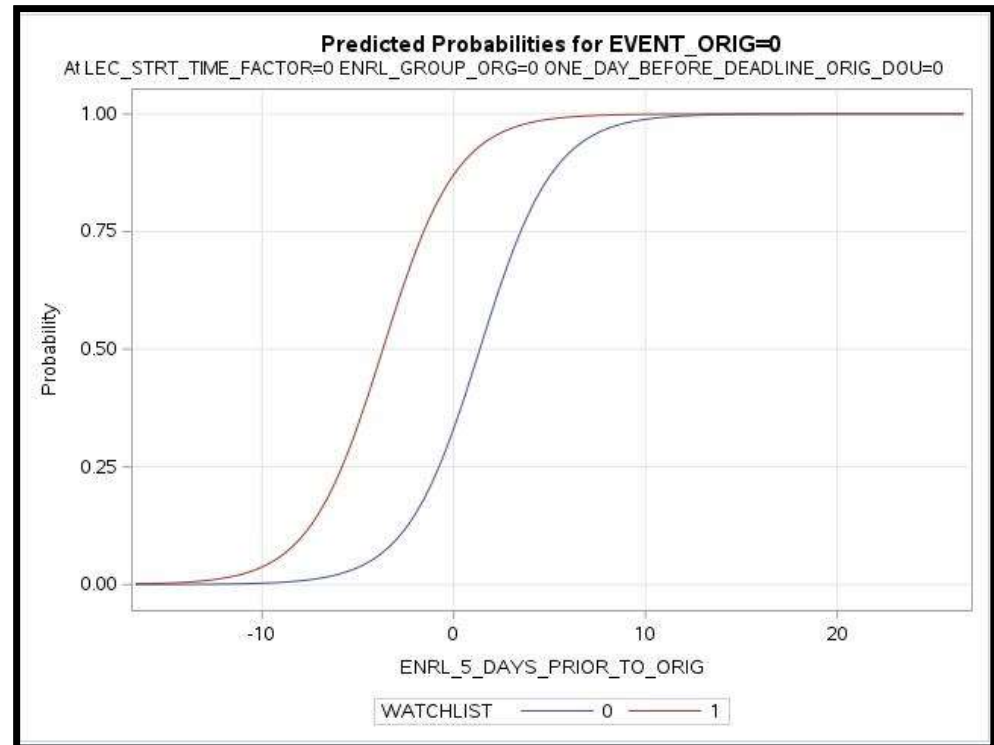




- We plot **ENRL\_5\_DAYS\_PRIOR\_TO\_ORIG** and **ONE\_DAY\_BEFORE\_DEADLINE\_ORIG\_DOUBLE\_DIGIT** into the same plot for a better visualization of OR.
- For each increment of enrollment during the five days period prior to 01/05/23, **sections with double digit enrollment** (coded as 1) have a higher chance of NOT being cancelled (**26.614** times) than those with single digit (coded as 0).



- We plot **ENRL\_5\_DAYS\_PRIOR\_TO\_ORIG** and **WATCHLIST** into the same plot for a better visualization of OR.
- For each increment of enrollment during the five days period prior to 01/05/23, **sections in the WATCHLIST** (coded as 1) have a higher chance of NOT being cancelled (**13.591** times) than those not being in the WATCHLIST (coded as 0).



# Extended Deadline (01/09/23) Variable Description

Variable Name	Variable Description
ENRL_5_DAYS_PRIOR_TO	The number of students who enrolled during the 5 days period prior to 01/09/23.
12.20.22_DOUBLE_DIGIT	Coded as 0 if enrollment on 12/20/22 for a section is a single digit (less than 10), and 1 for a double digit.
GROUP_ENRL_5_DAYS_PRIOR	Coded as 0 if the total enrollment 5 days prior to 01/09/23 is less than 5, and 1 if the total enrollment is $\geq 5$ .
STRT_DATE_ENRL_GROUP	Enrollment on 12/20/22, coded as 0 if enrollment is 0-5, coded as 1 if enrollment is 6-10, coded as 2 if enrollment is $>10$ .
ENRL_GROUP_EXT	Enrollment on 01/09/23, coded as 0 if enrollment is 0, coded as 1 if enrollment is 1-5, coded as 2 if enrollment is 6-10, coded as 3 if enrollment is $>10$ .
ONE_DAY_BEFORE_DEADLINE	Coded as 0 if enrollment on 01/08/23 for a section is a single digit (less than 10), and 1 for a double digit.
TIME_TO_CANCEL_EXT *	The number of days from 12/20/22 to the day a section is cancelled. If the section is not cancelled during this period, we assign the value 20. <b>*We create this variable to answer question 1 in page 6.</b>
HIGHLY_ENRL_EXT **	Coded as 1 if the average daily enrollment between 12/20/22 and 01/09/23 is $\geq 0.5$ , and 0 if the average is $< 0.5$ . <b>**We create this variable to answer question 2 in page 6.</b>
WATCHLIST	Coded as 1 if the section is on the watchlist, and 0 if not.
LEC_STRT_TIME_FACTOR	Coded as 0 if the lecture start time is between 8AM and 10AM, coded as 1 if the LEC start time is between 10AM and 12PM, coded as 2 if the LEC start time is between 12PM and 2PM, coded as 3 if the LEC start time is between 2PM and 4PM, coded as 4 if the LEC start time is after 4PM, coded as 5 for online classes.

# Model 2: Extended Deadline (01/09/23)

- In our analysis, EVENT\_EXT, our response (outcome) variable, is coded as:
  - **0**: The section is NOT cancelled by 01/09/23. This is our event of interest.
  - **1**: The section is cancelled by 01/09/23.
- After applying the backward stepwise selection technique, we obtain the final model that includes these significant factors (p-value < 0.05) at the 95% confidence level.

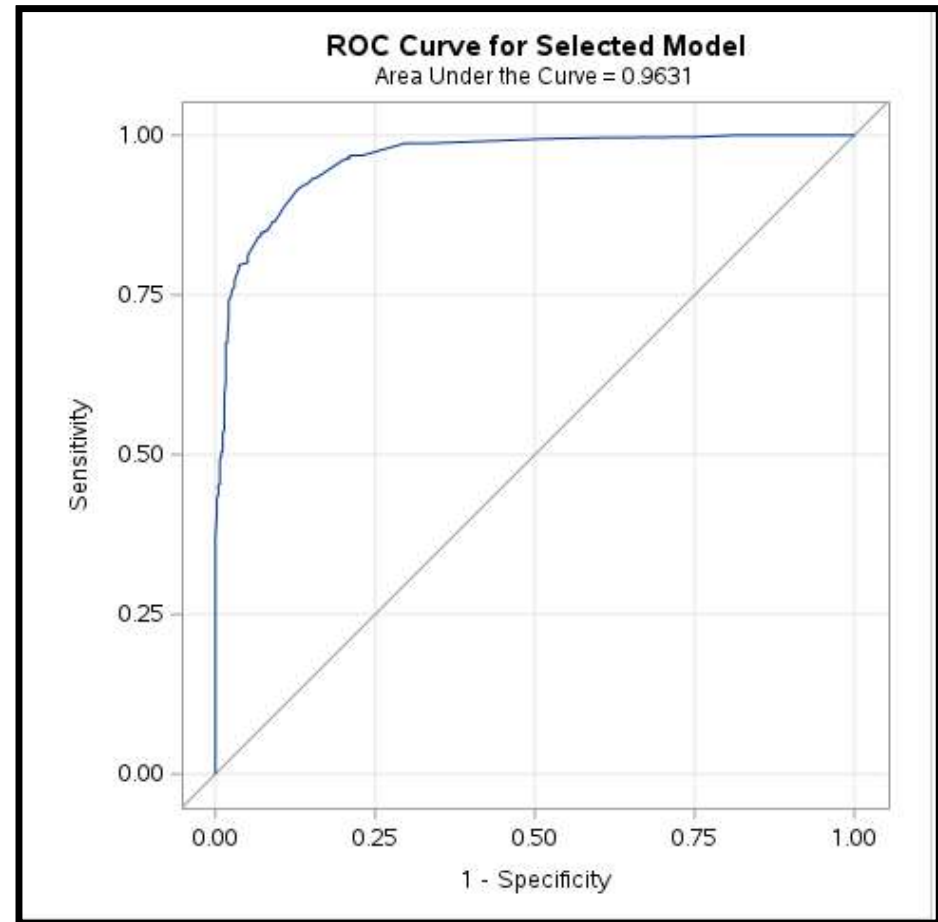
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3663	0.2586	83.7147	<.0001
ENRL_5_DAYS_PRIOR_TO		1	0.3819	0.1078	12.5486	0.0004
STRT_DATE_ENRL_GROUP	1	1	0.9642	0.3720	6.7180	0.0095
STRT_DATE_ENRL_GROUP	2	1	-0.9716	0.9227	1.1088	0.2923
ENRL_GROUP_EXT	1	1	3.0765	0.2198	195.8885	<.0001
ENRL_GROUP_EXT	2	1	4.9061	0.4011	149.6333	<.0001
ENRL_GROUP_EXT	3	1	8.3298	1.2767	42.5720	<.0001
WATCHLIST	1	1	2.0117	0.4498	20.0000	<.0001

Ordered Value	EVENT_EXT	Total Frequency
1	0	1587
2	1	434

Probability modeled is EVENT\_EXT='0'.

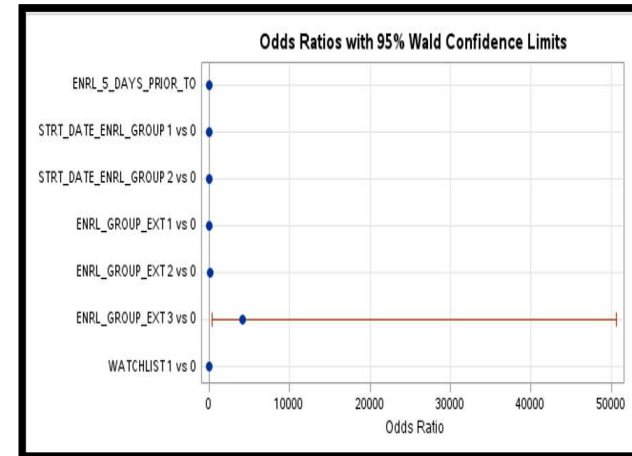
# Accuracy Level of Model 2

- The final model 2 has the AUC is 0.9631, which implies a **96.31%** chance that the model identifies which sections would be cancelled correctly.
- Let's interpret the impact of these significant factors in explaining the cancellation rate by using the **odds ratio**.

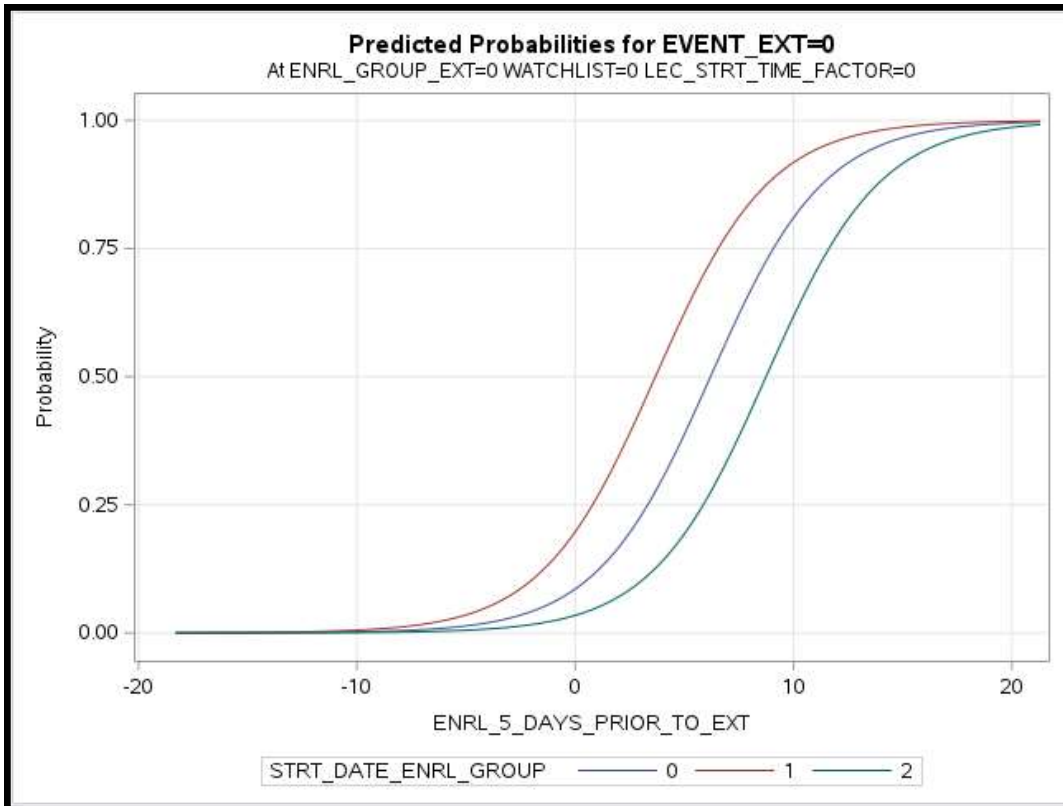


# Model 2 – Final Model's Odds Ratios

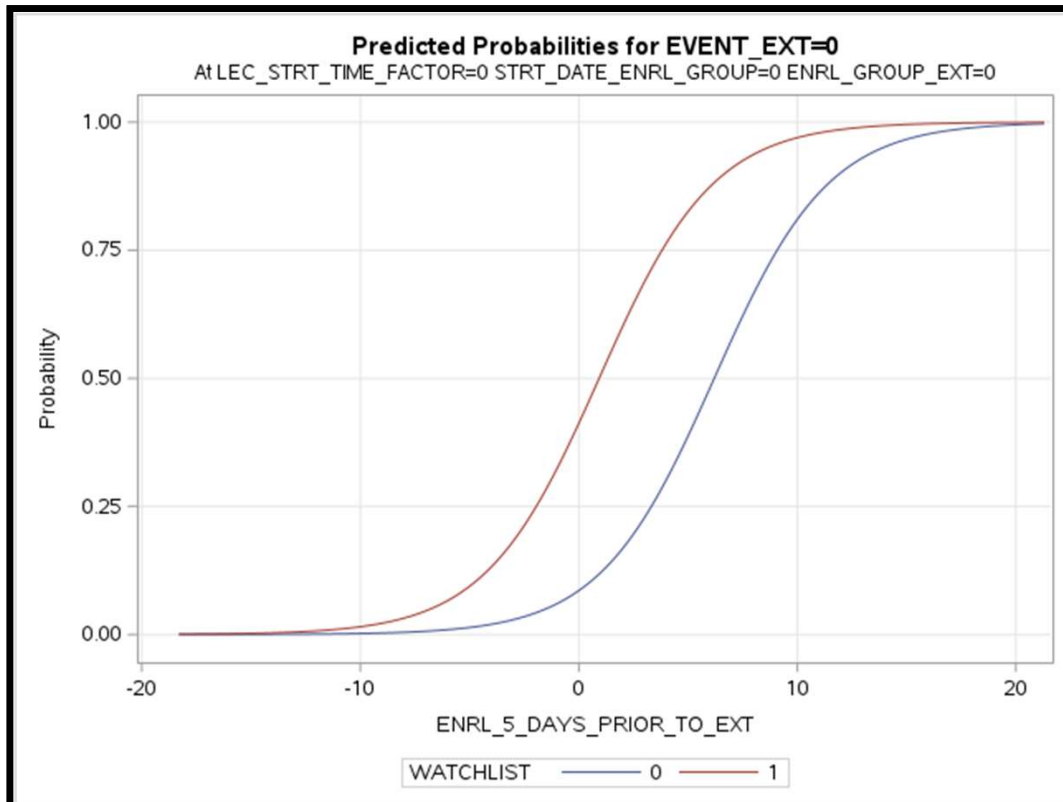
- Event of interest is NOT being cancelled by the Extended Deadline (01/09/23).
- Assuming other factors stay constant:
  - ENRL\_5\_DAYS\_PRIOR\_TO\_EXT has OR is **1.465**, which means for **each increment in enrollment during the 5 days period prior to 01/09/23**, the odds of NOT being cancelled of that section is **1.465 times** the odds of other sections.



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
ENRL_5_DAYS_PRIOR_TO	1.465	1.186	1.810
STRT_DATE_ENRL_GROUP 1 vs 0	2.623	1.265	5.438
STRT_DATE_ENRL_GROUP 2 vs 0	0.378	0.062	2.309
ENRL_GROUP_EXT 1 vs 0	21.683	14.093	33.359
ENRL_GROUP_EXT 2 vs 0	135.113	61.561	296.544
ENRL_GROUP_EXT 3 vs 0	>999.999	339.553	>999.999
WATCHLIST 1 vs 0	7.476	3.096	18.053

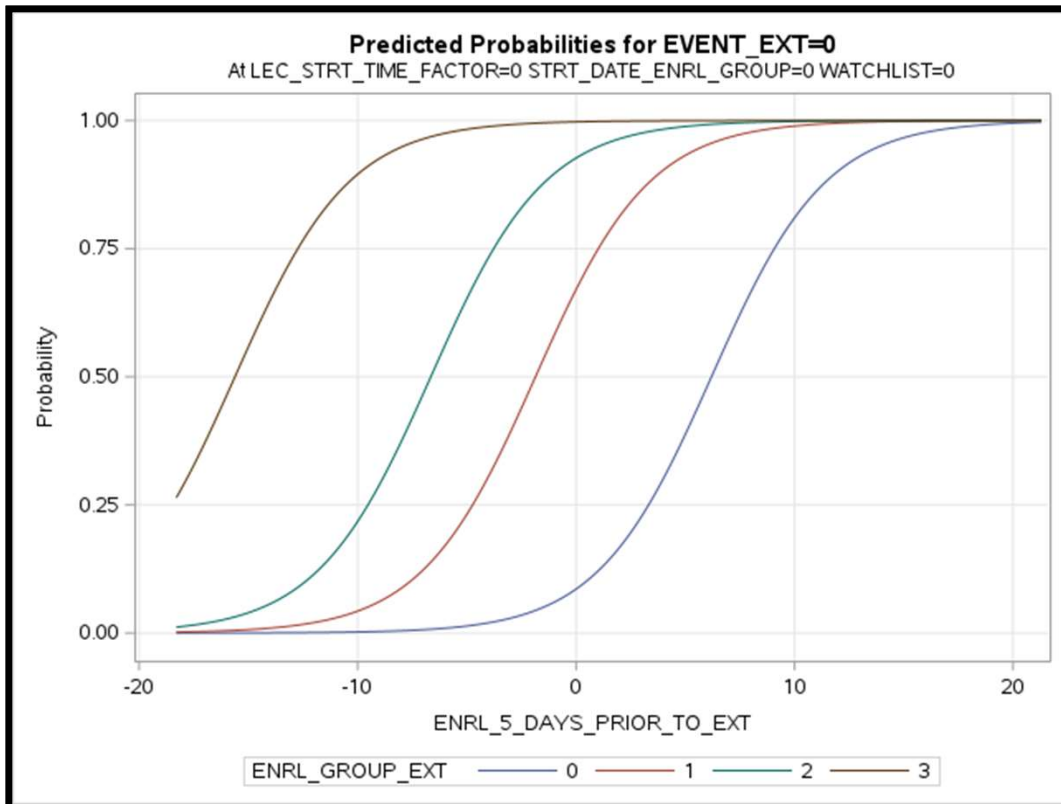


- We plot ENRL\_5\_DAYS\_PRIOR\_TO\_EXT and STRT\_DATE\_ENRL\_GROUP into the same plot for a better visualization of OR.
- For each increment of enrollment during the five days period prior to 01/09/23, **sections with 6-10 enrollment on 12/20/22** (coded as 1) have a higher chance of NOT being cancelled (**2.693** times) than **sections with 0-5 enrollment** (coded as 0).
- However, **sections with more than 10 enrollment** (coded as 2) have the odds of NOT being cancelled LOWER by **0.378** times than **sections with 0-5 enrollment**. This implies the **double - digit enrollment on 12/20/22** still could be at risk of being cancelled.



- We plot **ENRL\_5\_DAYS\_PRIOR\_TO\_EXT** and **WATCHLIST** into the same plot for a better visualization of OR.
- For each increment of enrollment during the five days period prior to 01/09/23, **sections in the watchlist** (coded as 1) have a higher chance of NOT being cancelled (**7.476 times**) than **sections not in the watchlist** (coded as 0).





- We plot **ENRL\_5\_DAYS\_PRIOR\_TO\_EXT** and **ENRL\_GROUP\_EXT** into the same plot for a better visualization of OR.
- For each increment of enrollment during the five days period prior to 01/09/23, **sections with more than 10 enrollment on 01/09/23** (coded as 3) have a higher chance of NOT being cancelled (more than **999.999 times**) than empty sections (coded as 0). The odds of NOT being cancelled for **the 6–10 enrollment group** (coded as 2) is **135.113**, and for **the 0–5 enrollment group** (coded as 1), the odds are **21.683** higher than those empty sections.

I just need  
the main ideas

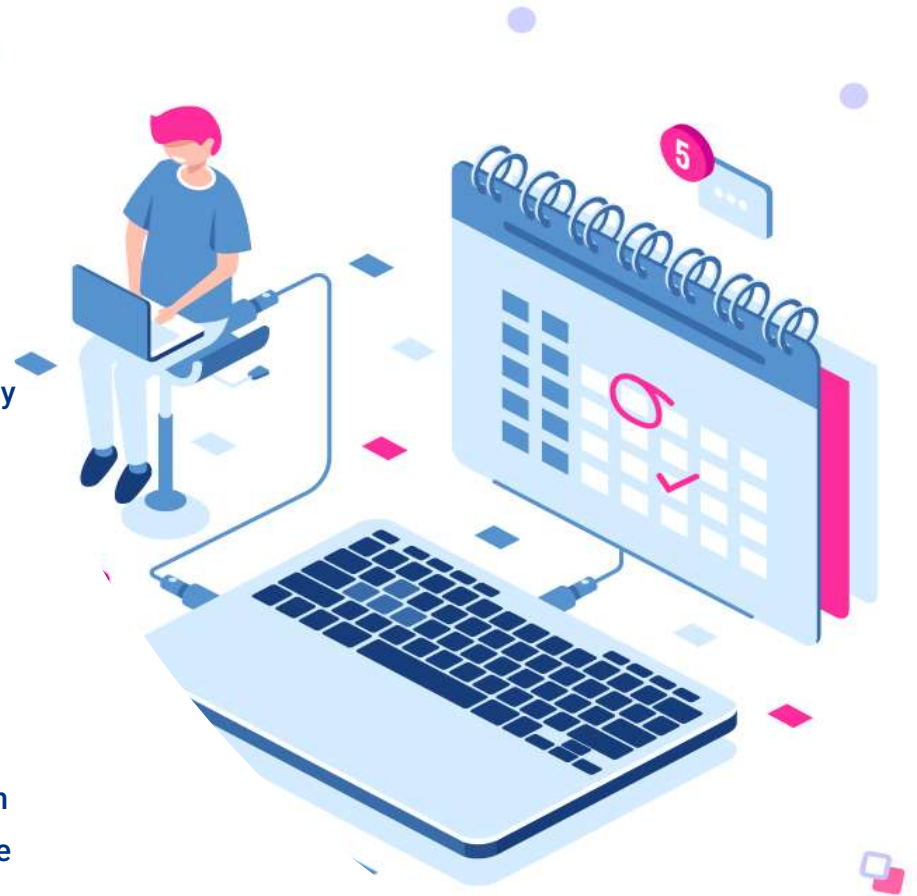


**Discussion and Conclusion**



# Discussion

- We created two logistic regression models (one for the original deadline – 01/05/23, and another one for the extended deadline – 01/09/23) to predict which sections will be cancelled (the accuracy rate is approximately 93% and 96%, respectively) and to answer three hypothesis questions on page 6.
1. **Does the extension period have a statistically significant impact on the number of cancelled sections?**
    - ✓ For both models, at 95% confidence level, there is NOT sufficient evidence to conclude TIME\_TO\_EVENT (the number of days from 12/20/22 to the day a section is cancelled) is statistically significant. In other words, extending the deadline does NOT significantly secure more sections from being cancelled.





# Discussion

## 2. Will highly-enrolled sections benefit from the extension period?

- ✓ Yes, but how we define “highly-enrolled” sections also play an important key here. We created a “highly-enrolled” variable (introduced in page 10 and 19), unfortunately, that variable is NOT significant in explaining the cancellation rate. Instead, both models show some indicators that we can consider “highly-enrolled”, or “lifesaving” (to be precise) for some sections:
  - ❖ “Enrollment during five day prior to the deadline” – the higher the enrollment, the better it is.
  - ❖ “Having a double-digit enrollment one day before the deadline”.
  - ❖ “Having more than 5 students enrolled by 6 AM on the deadline day”.
  - ❖ “Being in the watchlist”.

# Discussion



iconscout | Thai Hoang

id: 2294225

3. Which other factor(s) are statistically significant in explaining the cancellation rate?
  - ✓ In addition to the factors that we discussed in the previous page, the extended deadline model identifies the “Enrollment group on 12/20/22” as also being statistically significant in explaining the cancellation rate.
  - ✓ **Sections with 6–10 enrollment** have a higher chance of NOT being cancelled compared to **sections with 0–5 enrollment**.
  - ✓ But **sections with more than 10 enrollment** have a LOWER chance of NOT being cancelled compared to **sections with 0–5 enrollment**. Please keep in mind that all these sections are under enrolled on 12/20/22. This implies having high enrollment early on does not always mean that a section will make it by the cancellation deadline. During the more than two weeks period until the deadline, students can change their mind, drop/switch sections, and put those double-digit sections (on 12/20/22) at risk of being cancelled later.
  - ✓ One drawback of this variable is that we set the sections with **0-5 enrollment** as the reference to compare with other groups. Therefore, we cannot comment about the relationship between the two groups **6-10 enrollment** and **more than 10 enrollment**.

# Conclusion



**Five day prior to the deadline** is the “**prime period**” to determine whether the section can make it or not.



In addition, those significant factors (“**Having a double-digit enrollment one day before the deadline**”, “**Having more than 5 student enrolled by 6 AM on the deadline day**”) are “**lifesaving**” indicators which assist us to predict the odds of a section being cancelled.



Understanding the enrollment trend during the “**prime period**” and the “**lifesaving**” indicators assist the school of leadership to determine which sections should be added into the watchlist, which later may be beneficial in reducing their chance of cancellation.



According to this analysis, we recommend keeping the practice of allowing the school of leadership to make the decision to cancel certain sections on one day as well as add sections with the chance of surviving (that meet either some or all “**lifesaving**” indicators) into the watchlist and make a final decision on a later day.



In order to assist the school of leadership in making their data-driven decision for future enrollment monitoring periods, we offer two helpful tools:

- The [daily enrollment dashboard](#).
- Our predictive model to identify any section at risk of being cancelled.



## Contact Info

---

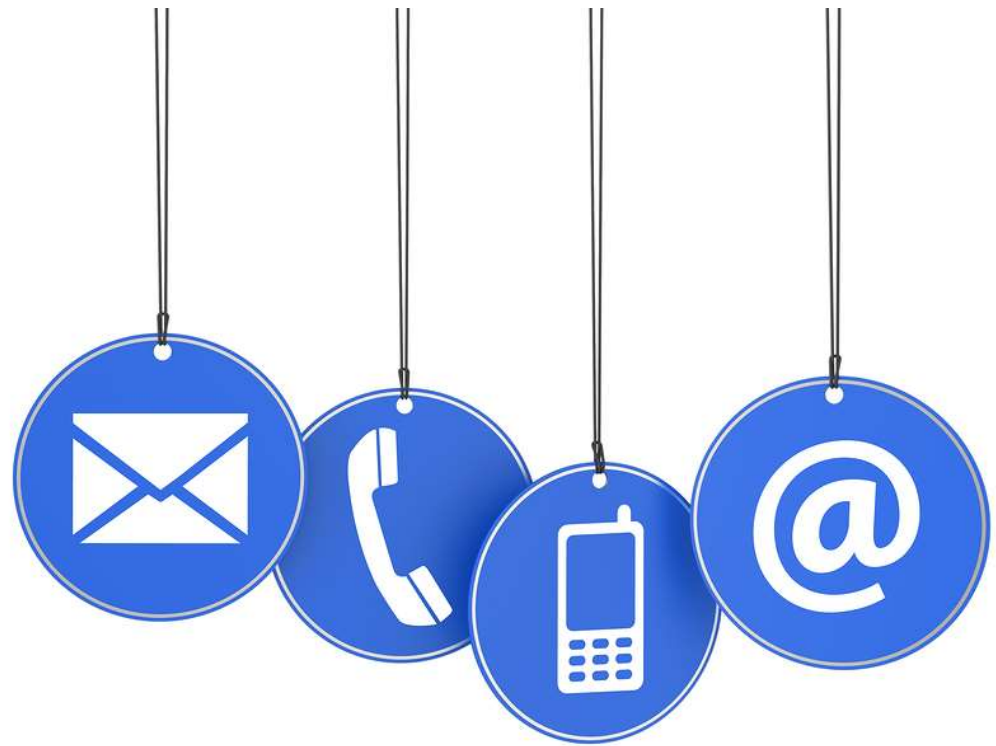
- Feel free to send your question/suggestion/discussion to:

Daniel Le

[dle@DallasCollege.edu](mailto:dle@DallasCollege.edu)

Dr. Jeremy Monteath-Valdez

[Jeremy.Valdez@DallasCollege.edu](mailto:Jeremy.Valdez@DallasCollege.edu)



**THANK YOU  
FOR LISTENING!**

