



Stakeholder Engagement: Responding to Questions about Your Predictive Modeling

A Rapid Insight Deep Dive

January 20, 2022

Meet Your Presenters



Wesley Pendarvis

*Senior Director, Edify
Partner Success*



James Cousins

Edify Product Analyst

Questions?

Throughout the presentation, feel free to submit questions using the Chat or Q&A.

Technical Difficulties?

Email RapidInsight@eab.com for assistance.

Submit a Question or Comment

The screenshot displays the EAB virtual session interface. At the top left is the EAB logo, consisting of a blue circle with a white building icon and the text "EAB". Below the logo, the text "EAB Virtual Session" is displayed in white and teal. The year "202" is partially visible. Two callout boxes with orange borders and yellow dots pointing to the interface provide instructions: "Use the Chat feature to send messages to all panelists or everyone" points to the "Chat" button, and "Use the Q&A feature to ask questions" points to the "Q&A" button. The bottom navigation bar includes "Audio Settings" with an upward arrow, "Chat", "Q&A", and "Leave Meeting" in red text.

EAB

EAB
Virtual Session

202

Use the Chat feature to send messages to all panelists or everyone

Use the Q&A feature to ask questions

Audio Settings ^

Chat

Q&A

Leave Meeting

- 1 **Rapid Insight and EAB**
- 2 Overview of the Presentation Structure
- 3 Common Transparency-Minded Questions
- 4 Audience Q & A

A New Partnership to Accelerate Data Democratization in Higher Education



Learn More:

Read more about EAB's latest partnership at
<https://eab.com/rapidinsight/>

We help schools support students from enrollment to graduation and beyond

➤ **ROOTED IN RESEARCH**

8,000+ Peer-tested best practices

500+ Enrollment innovations tested annually

➤ **ADVANTAGE OF SCALE**

2,100+ Institutions served

9.5 M+ Students supported by our SSMS

➤ **WE DELIVER RESULTS**

95% Of our partners continue with us year after year, reflecting the goals we **achieve together**

➤ Find and enroll your right-fit students

➤ Support and graduate more students



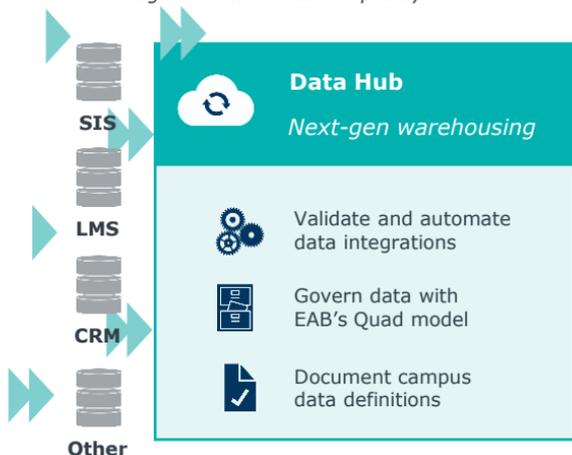
➤ Prepare your institution for the future

Inside the New and Improved Edify

An Education Data Platform to Accelerate Campus Data Strategy

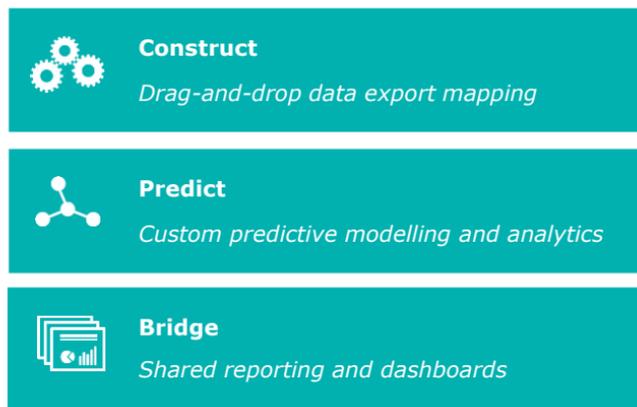
Centralized Data

Integrate data across campus systems



Decentralized Analytics

Activate your professional and citizen data scientists



EAB Professional and Technical Data Services

Strategy and
Culture Support

Integration
Services

Process
Consulting

Data Science
and Reports

- 1 Rapid Insight and EAB
- 2 **Overview of the Presentation Structure**
- 3 Common Transparency-Minded Questions
- 4 Audience Q & A

Overview of the Presentation Structure



- 1 Rapid Insight and EAB
- 2 Overview of the Presentation Structure
- 3 **Common Transparency-Minded Questions**
- 4 Audience Q & A

Common Transparency-Minded Questions



- 1 How did you choose the variables in the model?
- 2 Did you consider {variable name}?
- 3 How do you know {variable name} is (or is not) related?
- 4 Are you omitting restricted characteristics?
- 5 How do you know the model is not overfitting?

Question 1

How did you choose the variables in the model?



First, Pinpoint the Question's Origin

Is the Question About Data Preparation or Modeling Methodology?

Two Potential Origins

1



Statistically

Why did these specific variables from your dataset enter your model?

2



Practically

Why did you choose these variables for your dataset?

Explain Inclusion of Statistically Significant Variables

Why did these specific variables from your dataset enter your model?

Highest-Level Response:

“Automatic statistical mining ruled out characteristics not statistically related to the outcome, and left in those with significant relationships.”

The screenshot shows a software interface for model building. At the top, there is a navigation bar with icons for Statistics, View Data, Visualize, Correlation, Analyze, Clustering, and Model. Below this is the 'Analysis Info' section, which displays 'Total Records in DataSet: 5618' and 'Variables in DataSet: 27'. It also shows 'Records Using: 5618' and 'Variables in Analysis: 27'. The 'Y Variable' is set to 'Enroll'. There are tabs for 'Automine' and 'Model Building'. A 'P-Value' dropdown is set to '0.01' and another dropdown is set to 'Automine'. Below this, there are options for 'Set Model Availability', 'Set Missing Handling', and a summary: 'Related: 13 Unrelated: 8 Auto-created variables: 32'. A table lists the variables included in the model:

Name	Type	Created By	F Value	Correlation	Model Availability
Applied for FinAid	Binary	Data Source	0.000020	-0.000060	Available if Related to Y
Attended Event w Family Members	Binary	Data Source	10.81	0.04382	Available if Related to Y
Citizenship Code	Continuous	Data Source	1.397	-0.01577	Available if Related to Y
Days Between App and Term Start	Continuous	Data Source	97.90	-0.1309	Available if Related to Y
Department	Categorical	Data Source	8.832	0.04838	Available if Related to Y

Inclusion of Statistically Significant Variables continued

Why did these specific variables from your dataset enter your model?

Lower-Level Response:

“From among the significant factors, the most explanatory field is identified, then added. This continues until nothing left improves the model.”

Model Steps

- Step #6
- Step #7
- Step #8
- Step #9
- Step #10

Candidate Variables	Score
Days Between App and Term Start	70.55
Binary(Ethnicity,White, non hispanic)	36.48
Binary(Ethnicity,African-American)	35.60
LOGe(Distance from Campus)	21.13
Legacy	6.405

Variable entered = Days Between App and Term Start

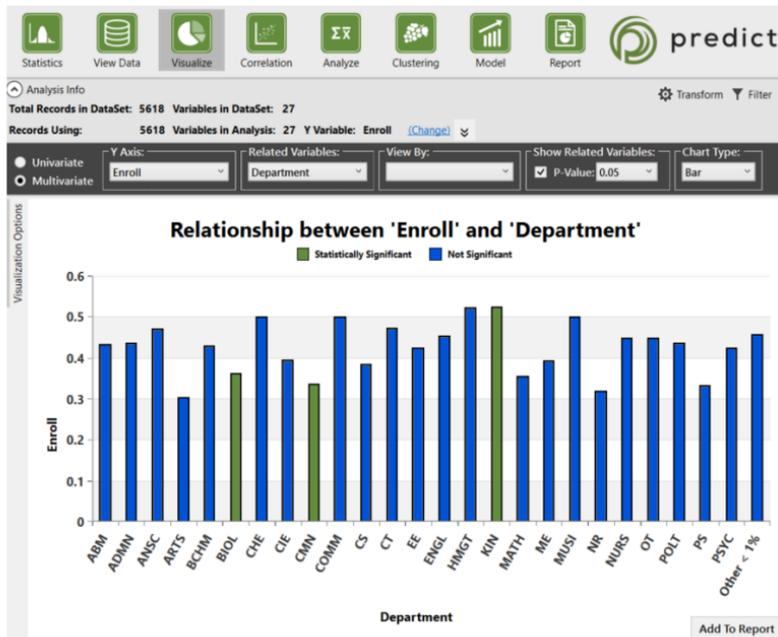
Model Steps Final Regression Model

Identify Opportunities for Collaborative Discussion

Why did you choose these variables for your dataset?

For instance, does the stakeholder:

- Have concerns about any of the variables?
- Have suggestions for other fields?
- Want to know more about where the data comes from?



Question 2

Did you consider {variable name}?

Answers at Multiple Levels of Detail - 1st Test

The screenshot shows a data analysis tool interface. At the top, there are icons for Statistics, View Data, Visualize, Correlation, Analyze, and Clustering. Below these is an 'Analysis Info' section with the following details:

- Total Records in DataSet: 5618
- Variables in DataSet: 27
- Records Using: 5618
- Variables in Analysis: 27
- Y Variable: Enroll

A 'Columns' list on the left shows various variables with checkboxes. A table on the right lists the variables and their types:

Variable	Type
Inst_Need_Grant	Categorical
SAT_Verbal	Continuous
Citizenship Code	Continuous
Nationality	Categorical
Department	Categorical
First Generation	Binary
Web Applicant	Binary
Admitted	Constant
ID	Text
Application Date	Date
Term Start Date	Date
Attended Event w Family Members	Binary
Term	Continuous

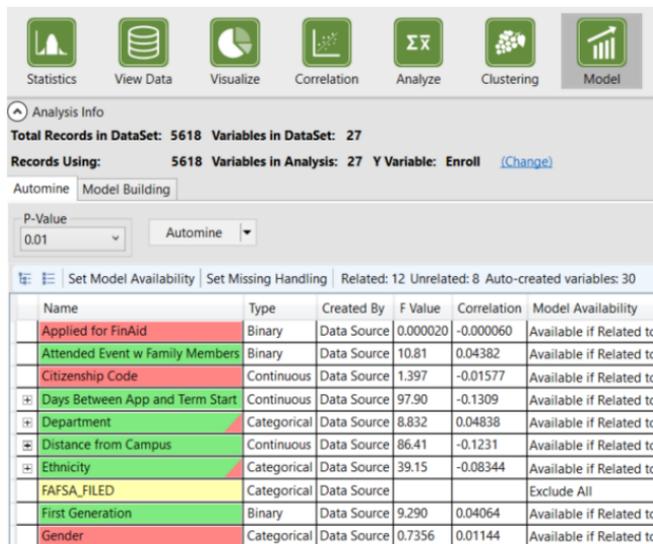
If you don't see the variable included in your dataset, Predict is not "considering" that field as a possible predictor.

Further, if the variable is

- "Constant"
- "Date" or
- "Text"

Predict cannot test or use that variable.

Answers at Multiple Levels of Detail - 2nd Test



Statistics View Data Visualize Correlation Analyze Clustering Model

Analysis Info

Total Records in DataSet: 5618 Variables in DataSet: 27

Records Using: 5618 Variables in Analysis: 27 Y Variable: Enroll (Change)

Automine Model Building

P-Value: 0.01 Automine

Set Model Availability Set Missing Handling Related: 12 Unrelated: 8 Auto-created variables: 30

Name	Type	Created By	F Value	Correlation	Model Availability
Applied for FinAid	Binary	Data Source	0.000020	-0.000060	Available if Related to
Attended Event w Family Members	Binary	Data Source	10.81	0.04382	Available if Related to
Citizenship Code	Continuous	Data Source	1.397	-0.01577	Available if Related to
Days Between App and Term Start	Continuous	Data Source	97.90	-0.1309	Available if Related to
Department	Categorical	Data Source	8.832	0.04838	Available if Related to
Distance from Campus	Continuous	Data Source	86.41	-0.1231	Available if Related to
Ethnicity	Categorical	Data Source	39.15	-0.08344	Available if Related to
FAFSA_FILED	Categorical	Data Source			Exclude All
First Generation	Binary	Data Source	9.290	0.04064	Available if Related to
Gender	Categorical	Data Source	0.7356	0.01144	Available if Related to

Not Eligible for the Model

Red Shading
Not significant at the specified p-value

Yellow Shading
Excluded by user

Eligible for the Model

Green Shading
Significant at the specified p-value

Green Shading w/ Accent
Automatically created transformations are related

Answers at Multiple Levels of Detail - 3rd Test



Statistics View Data Visualize Correlation Analyze Clustering Model Report

Analysis Info
Total Records in DataSet: 5618 Variables in DataSet: 27
Records Using: 5618 Variables in Analysis: 27 Y Variable: Enroll (Change)

Automine Model Building

Variables:
Attended Event w Family Members
Distance from Campus
First Generation
Legacy
SAT Math
SAT_Verbal

Included Variables:
Binary(Ethnicity,White, non hispanic)
Binary(Inst_Need_Grant,0)
Days Between App and Term Start
In_State
LOGe(Days Between App and Term Start)
LOGe(Distance from Campus)
LOGe(SAT Math)
SAT Comp

Build
Suggest Variable
Build Stepwise
Build Automatically

View 'new variable' suggestions

Variables

The variables that were available for inclusion in the model but did not get included.

Included Variables

The variables that entered and remained in the model.

Use What You Know!



Questions and answers are all situated inside of relationships. Use what you know about the person who's asking to answer at the level they're seeking.

Question 3

How do you know {variable name}
is (or is not) related?

Leverage the Categorized List of Variables



Statistics
View Data
Visualize
Correlation
Analyze
Clustering
Model

Analysis Info

Total Records in DataSet: 5618 Variables in DataSet: 27

Records Using: 5618 Variables in Analysis: 27 Y Variable: Enroll [\(Change\)](#)

Automine Model Building

P-Value: 0.01 Automine Automine

[Set Model Availability](#) | [Set Missing Handling](#) | Related: 12 Unrelated: 8 Auto-created variables: 30

Name	Type	Created By	F Value	Correlation	Model Availability
Applied for FinAid	Binary	Data Source	0.000020	-0.000060	Available if Related to
Attended Event w Family Members	Binary	Data Source	10.81	0.04382	Available if Related to
Citizenship Code	Continuous	Data Source	1.397	-0.01577	Available if Related to
Days Between App and Term Start	Continuous	Data Source	97.90	-0.1309	Available if Related to
Department	Categorical	Data Source	8.832	0.04838	Available if Related to
Distance from Campus	Continuous	Data Source	86.41	-0.1231	Available if Related to
Ethnicity	Categorical	Data Source	39.15	-0.08344	Available if Related to
FAFSA_FILED	Categorical	Data Source			Exclude All
First Generation	Binary	Data Source	9.290	0.04064	Available if Related to
Gender	Categorical	Data Source	0.7356	0.01144	Available if Related to

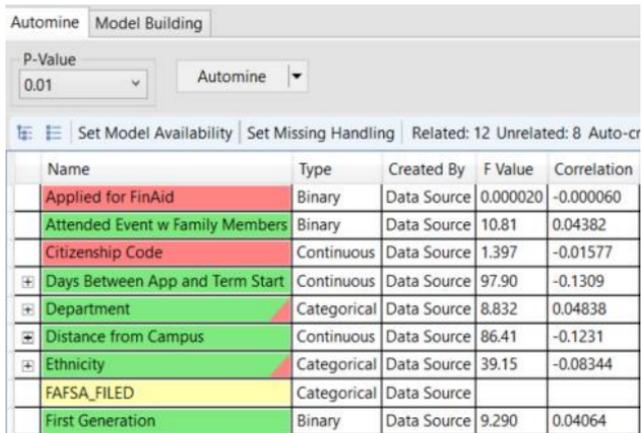
P-Value
The likelihood that a pattern will randomly occur

Red Shading
Not significant at the specified p-value

Green Shading
Significant at the specified p-value

Green Shading w/ Accent
Automatically created transformations are related

Spark a New Conversation



The screenshot shows the Automine Model Building interface. At the top, there are tabs for 'Automine' and 'Model Building'. Below the tabs, there is a 'P-Value' dropdown set to '0.01' and a 'Automine' dropdown. Below that, there are buttons for 'Set Model Availability', 'Set Missing Handling', and a status bar showing 'Related: 12 Unrelated: 8 Auto-cr'. The main part of the interface is a table with the following columns: Name, Type, Created By, F Value, and Correlation.

Name	Type	Created By	F Value	Correlation
Applied for FinAid	Binary	Data Source	0.000020	-0.000060
Attended Event w Family Members	Binary	Data Source	10.81	0.04382
Citizenship Code	Continuous	Data Source	1.397	-0.01577
Days Between App and Term Start	Continuous	Data Source	97.90	-0.1309
Department	Categorical	Data Source	8.832	0.04838
Distance from Campus	Continuous	Data Source	86.41	-0.1231
Ethnicity	Categorical	Data Source	39.15	-0.08344
FAFSA_FILED	Categorical	Data Source		
First Generation	Binary	Data Source	9.290	0.04064

from:

“How do you know international status isn’t related?”



to:

“Oh, how does First Generation impact enrollment?”

Question 4

Are you omitting
restricted characteristics?

Describe the Process for Omitting Variables



Automine Model Building

P-Value: 0.01 Automine

Set Model Availability | Set Missing Handling | Related: 12 Unrelated: 8 Auto-created variables: 30

Name	Type	Created By	F Value	Correlation	Model Availability
Applied for FinAid	Binary	Data Source	0.000020	-0.000060	Available if Related to Y
Attended Event w Family Members	Binary	Data Source	10.81	0.04382	Available if Related to Y
Citizenship Code	Continuous	Data Source	1.397	-0.01577	Available if Related to Y
Days Between App and Term Start	Continuous	Data Source	97.90	-0.1309	Available if Related to Y
Department	Categorical	Data Source	8.832	0.04838	Available if Related to Y
Distance from Campus	Continuous	Data Source	86.41	-0.1231	Available if Related to Y
Ethnicity	Categorical	Data Source	39.15	-0.08344	Available if Related to Y
FAFSA FILED					All
First Generation					if Related to Y
Gender					if Related to Y
In_State					if Related to Y
Inst_Need_Grant					if Related to Y
Legacy	Binary	Data Source	36.23	0.08006	Available if Related to Y
Nationality	Categorical	Data Source	5.893	0.03238	Available if Related to Y
RI_totFamilies_T_2	Categorical	Data Source	13.97	0.07524	Available if Related to Y
SAT Comp	Continuous	Data Source	87.39	-0.1256	Available if Related to Y
SAT Math	Continuous	Data Source	33.33	-0.07797	Available if Related to Y
SAT_Verbal	Continuous	Data Source	165.33	-0.1716	Available if Related to Y
Term	Continuous	Data Source	0.000020	0.000060	Available if Related to Y
Urban_Rural_Indicator	Categorical	Data Source	3.030	0.02322	Available if Related to Y
Web Applicant	Binary	Data Source	0.8128	0.01203	Available if Related to Y

Context menu options:

- Make Variables and their transforms available for modeling
- Exclude Variables and their transforms from modeling
- Set Variables and their transforms' Missing Handlings to...
- Export To .CSV

“Excluding” fields in the automine tab allows you to leave variables in your overall analysis (for descriptive purposes) but ensures it does not enter your model

Question 5

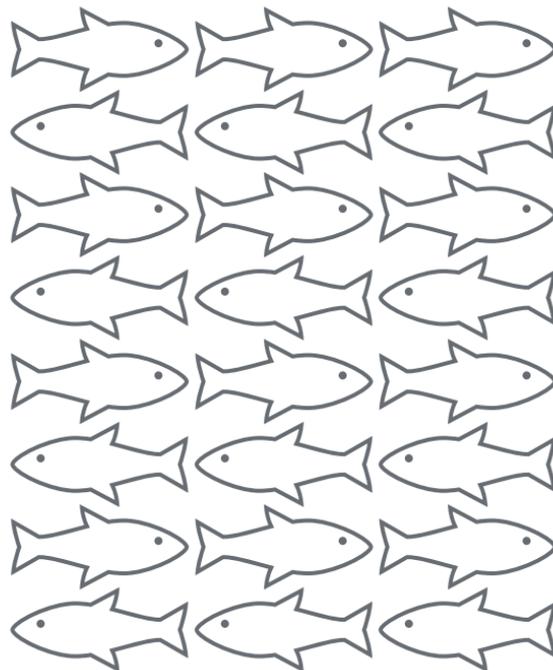
How do you know the model
is not overfitting?

Defining What It Means to Overfit a Model



Overfitting is the process of building a model which relies too heavily on a training population's behaviors.

You can also think of it as including more variables than appropriately "fit" in a model.



Check If You Have Overfitted a Model

Overfitting *starts* where statistical significance *stops*

Final Regression Model

Predicting: Enroll

Variable	Coef	S.E.	Wald chi-sqr	p-value
Intercept	-507.58	27.72	335.30	2.0006e-13
Binary(Ethnicity,White, non hispanic)	0.3515	0.1300	7.312	0.00685
Binary(Inst_Need_Grant,0)	-0.5209	0.1406	13.72	0.000212
Days Between App and Term Start	0.04220	0.00959	19.35	0.000011
In_State	0.8032	0.1237	42.13	2.0006e-13
Legacy	0.07177	0.1812	0.1568	0.6921
LOGe(Days Between App and Term Start)	-11.58	2.312	25.09	5.4822e-7
LOGe(Distance from Campus)	-0.1937	0.06087	10.13	0.00146
LOGe(SAT Math)	106.45	5.022	449.41	2.0006e-13
SAT Comp	-0.09884	0.00463	456.50	2.0006e-13

In order to build a model which overfits (in Predict), users must manually add fields themselves.

- Significance testing “checks” to see if a pattern of behavior is systematic enough that it’s likely to occur in an upcoming population.
- The fact that **legacy** has a p-value above 0.05 (or the threshold being used at the time) indicates this model is “overfit” to this training population.

- 1 Rapid Insight and EAB
- 2 Overview of the Presentation Structure
- 3 Common Transparency-Minded Questions
- 4 Audience Q & A

Q & A: Submit a Question Using the Chat



Wesley Pendarvis

*Senior Director, Edify
Partner Success*



James Cousins

Edify Product Analyst



Additional Questions?

Email RapidInsight@eab.com to
chat with our data experts directly.

Quick Poll

How was today's session?

Please take a few minutes to complete the survey to provide additional feedback!



Washington DC | Richmond | Birmingham | Minneapolis | New York | Chicago

202-747-1000 | eab.com