EAB

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

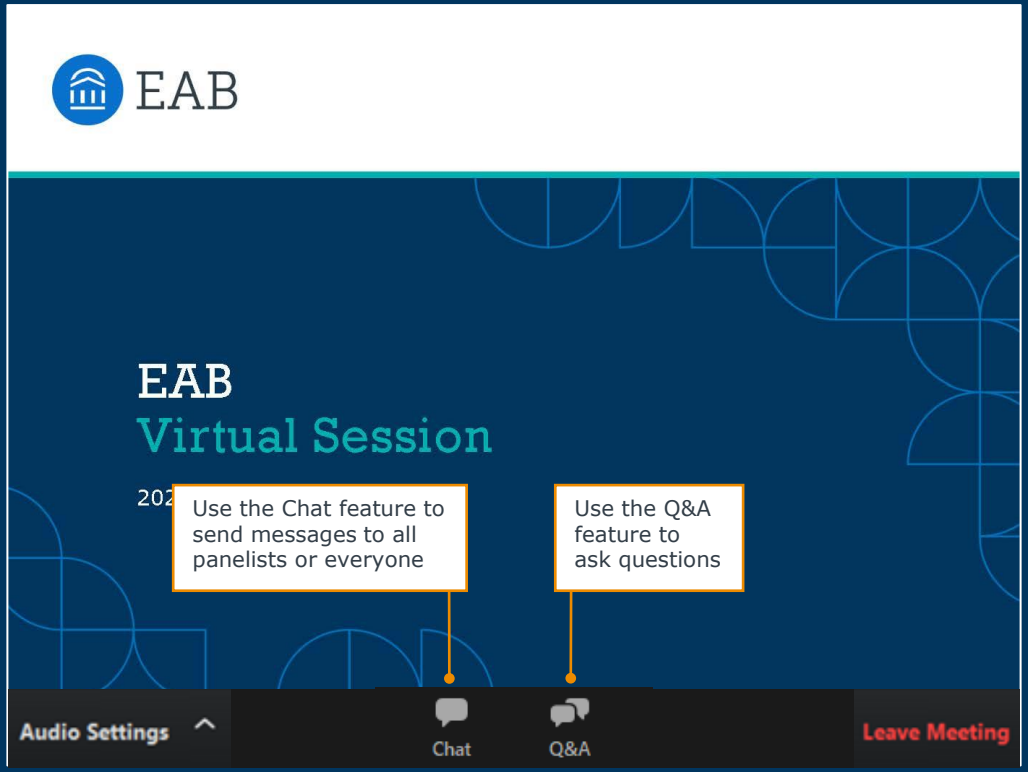
Questions?

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



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/

Source: EAB interviews and analysis.

EAB

P-20 | Community Colleges | Four-Year Colleges and Universities | Graduate and Adult Learning

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**

Inside the New and Improved **Edify**

An Education Data Platform to Accelerate Campus Data Strategy

Centralized Data

Integrate data across campus systems

Data Hub

Next-gen warehousing

- Validate and automate data integrations
- Govern data with EAB's Quad model
- Document campus data definitions

Decentralized Analytics

Activate your professional and citizen data scientists

Construct

Drag-and-drop data export mapping

Predict

Custom predictive modelling and analytics

Bridge

Shared reporting and dashboards

EAB Professional and Technical Data Services

- Strategy and Culture Support
- Integration Services
- Process Consulting
- Data Science and Reports

Source: Edify, EAB's Education Data Platform.

1 Rapid Insight and EAB

2 **Overview of the Presentation Structure**

3 Common Transparency-Minded Questions

4 Audience Q & A

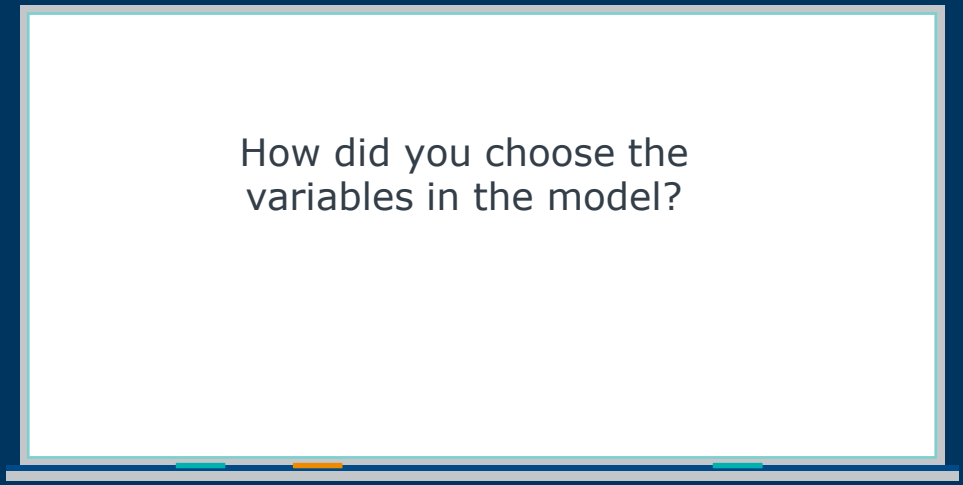
Overview of the Presentation Structure

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



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

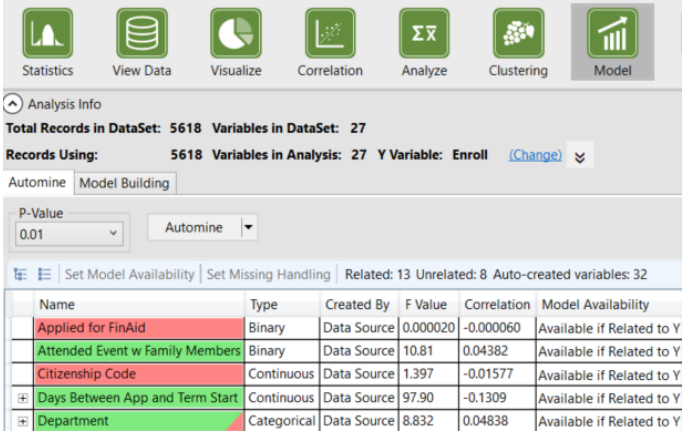
Why did you choose these variables for your dataset?

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."

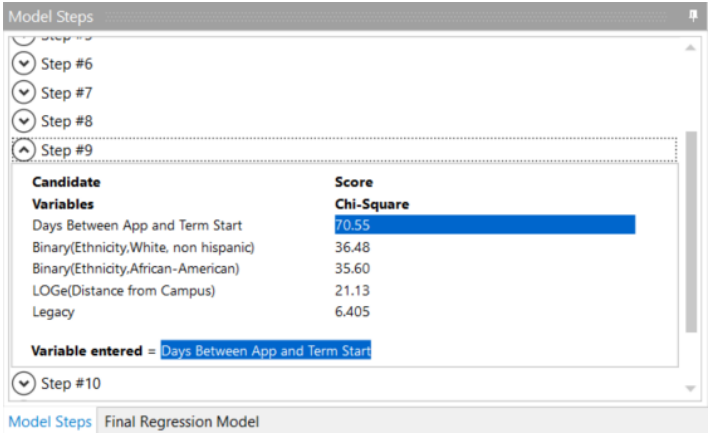


Inclusion of Statistically Significant Variables continued

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."

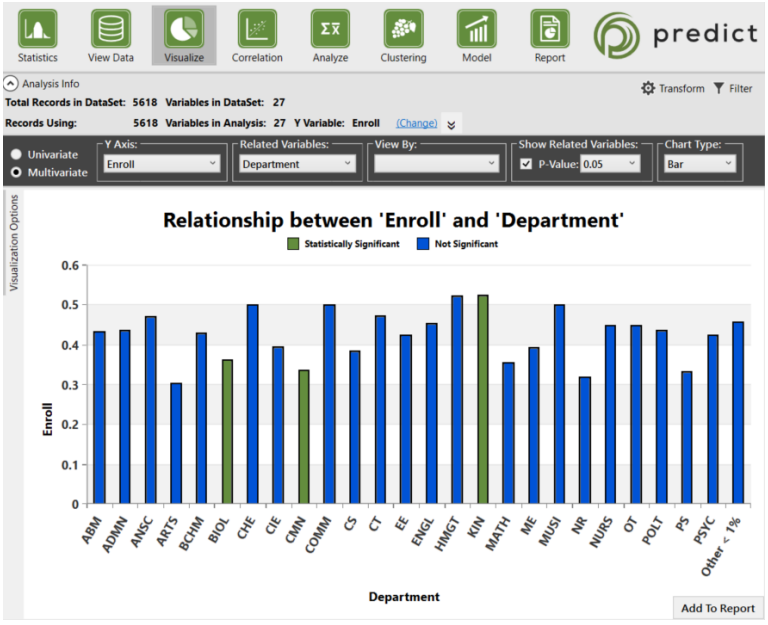


Identify Opportunities for Collaborative Discussion

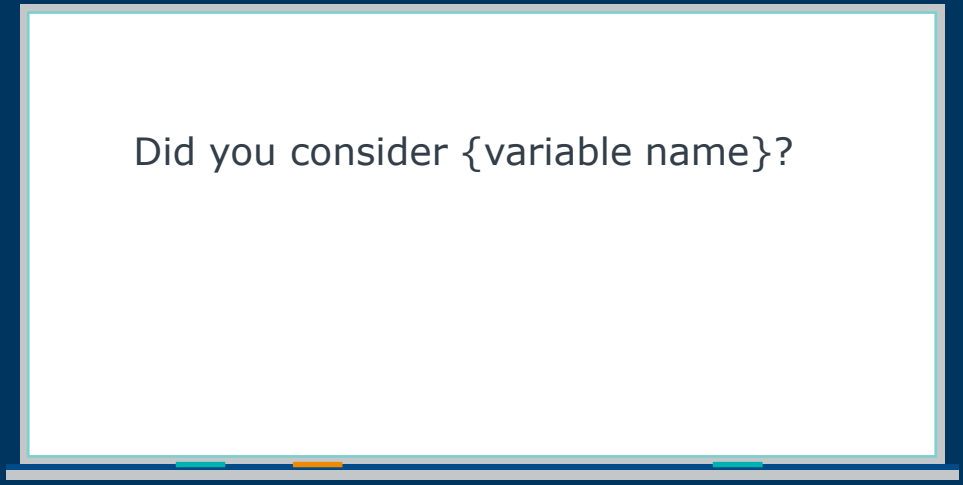
Why did you choose these variables for your dataset?

For instance, does the stakeholder:

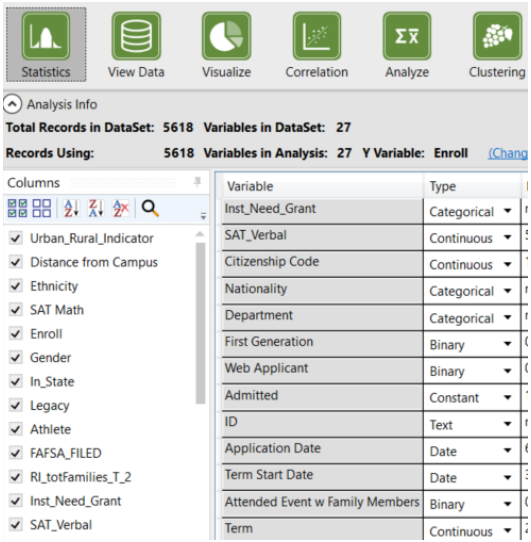
- Have concerns about any of the variables?
- Have suggestions for other fields?
- Want to know more about where the data comes from?



Question 2



Answers at Multiple Levels of Detail - 1st Test



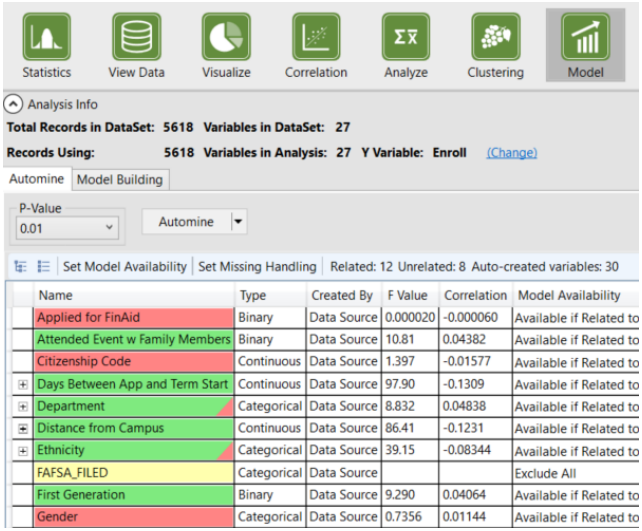
If you don't see the variable included in your dataset, Predict is not "considering" that field as a possible predictor.

Further, if the variable is

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

Predict cannot test or use that variable.

Answers at Multiple Levels of Detail - 2nd Test



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

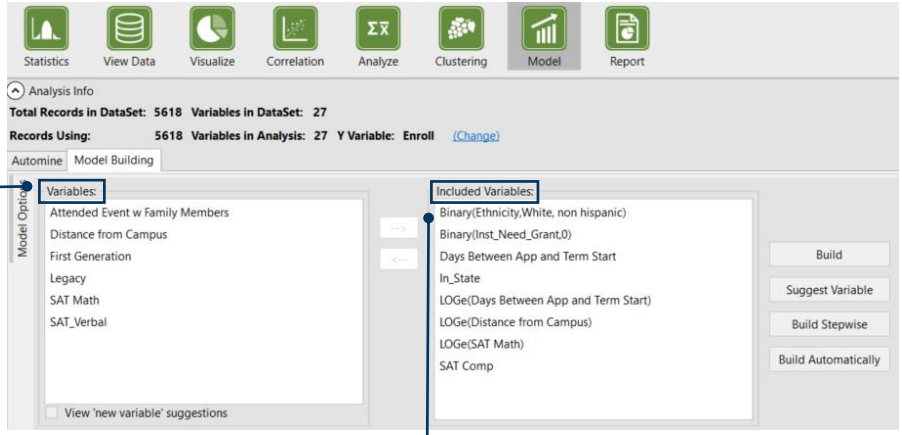
Green Shading
Significant at the specified p-value

Green Shading w/ Accent
Automatically created transformations are related

Answers at Multiple Levels of Detail - 3rd Test

Variables

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



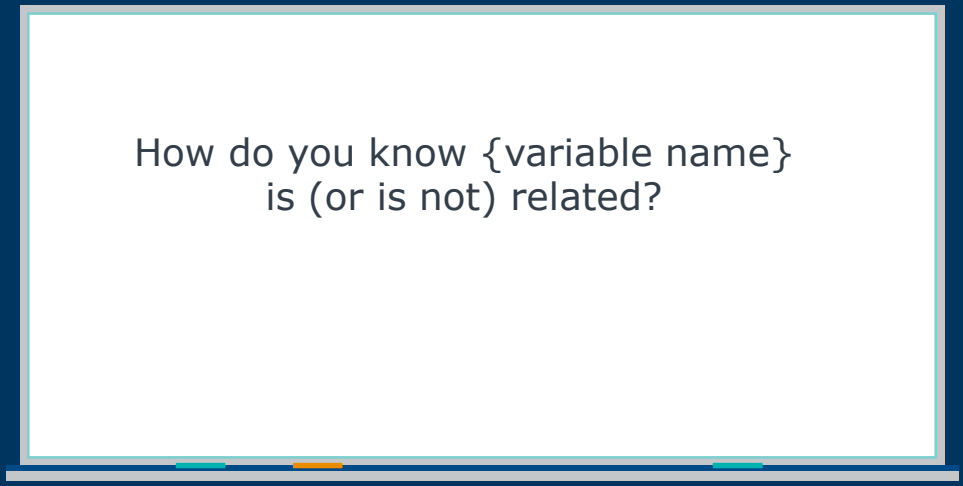
Included Variables

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



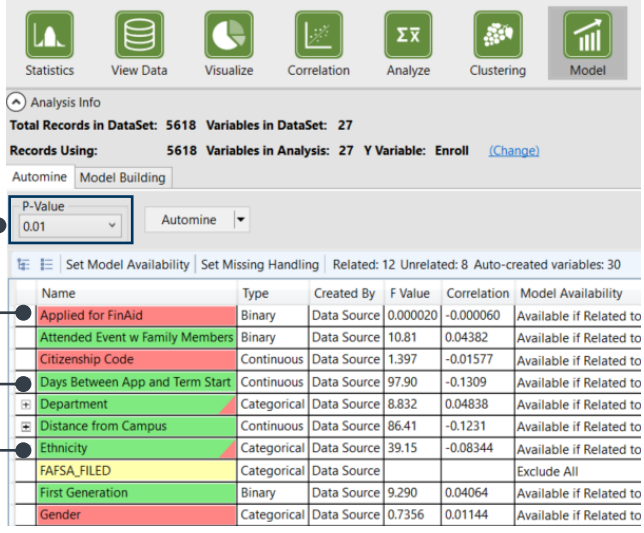
Leverage the Categorized List of Variables

P-Value
The likelihood that a pattern will randomly occur

Red Shading
Not significant at the specified p-value

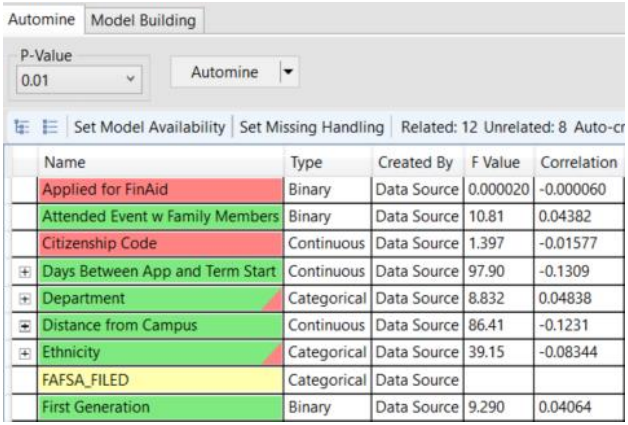
Green Shading
Significant at the specified p-value

Green Shading w/ Accent
Automatically created transformations are related



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

Spark a New Conversation



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

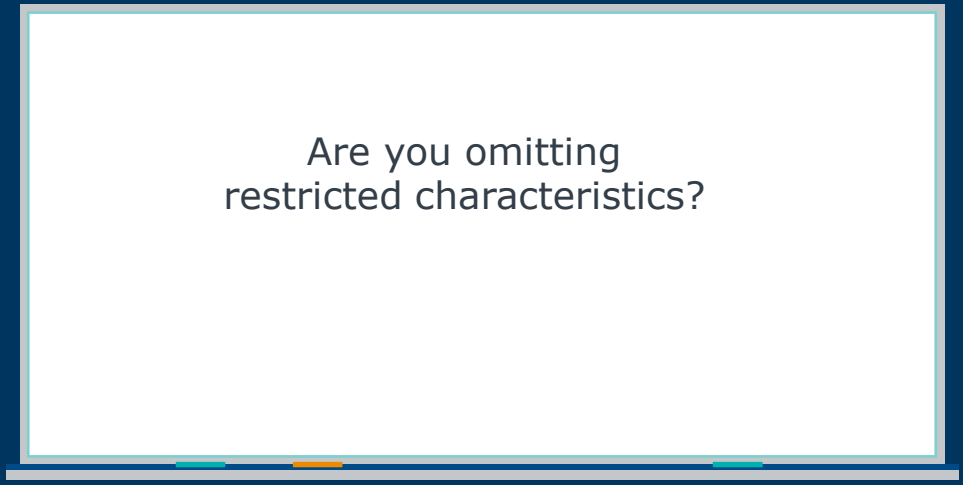
from:

"How do you know international status isn't related?"

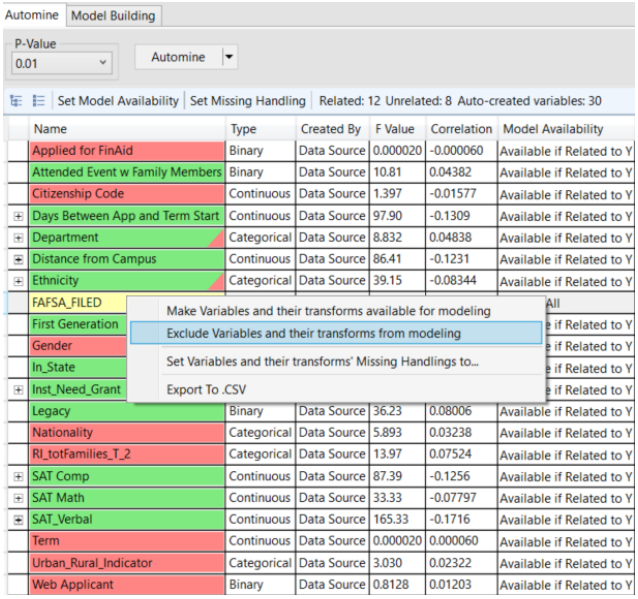
to:

"Oh, how does First Generation impact enrollment?

Question 4



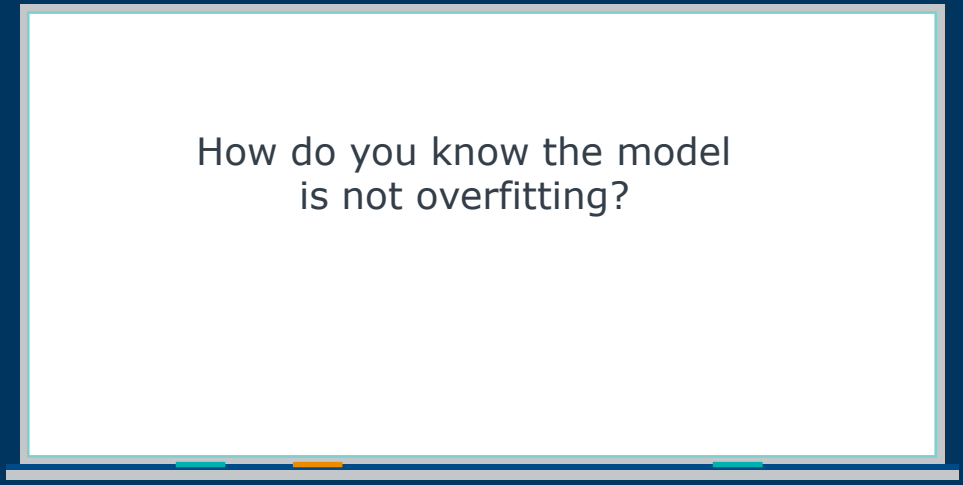
Describe the Process for Omitting Variables



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					e if Related to Y
Gender					e if Related to Y
In_State					e if Related to Y
Inst_Need_Grant					e 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

"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

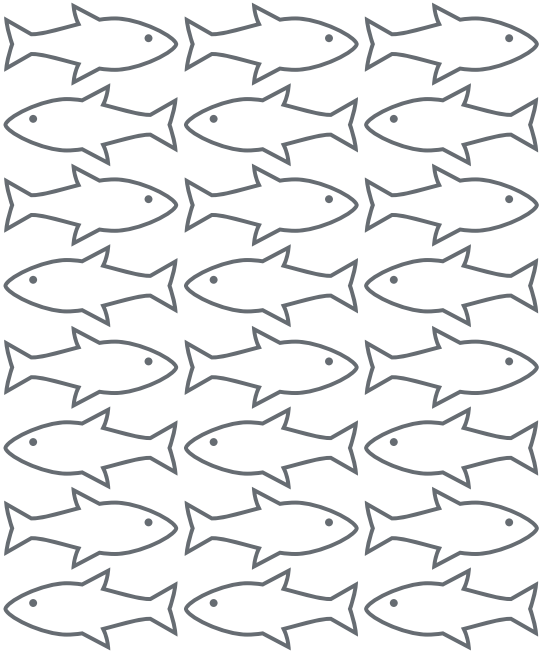
Question 5



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