12 IDEAS THAT WILL TRANSFORM YOUR PREDICTIVE MODELS

FINAL PRESENTATION FROM THE 2023-2024 BODILY BICENTENNIAL PROFESSOR IN ANALYTICS, UVA DARDEN SCHOOL OF BUSINESS

DEAN ABBOTT, ABBOTT ANALYTICS

HTTPS://WWW.LINKEDIN.COM/IN/DEANABBOTT/

@DEANABB

THANK YOU!

Image Built with DALL-E

WHY THESE “IDEAS” MATTER: THE MINDSET OF THE INDUSTRY

- Key Ideas in Machine Learning Literature
  - Algorithms
  - Math

WHY THESE "IDEAS" MATTER: THE MINDSET OF THE INDUSTRY

- Key Ideas in Machine Learning Literature
  - Algorithms
  - Math
- Representative of Industry Mindset: Top 2 most popular books in Master Data Science programs
  - **#1:** Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009.
    #2: Bishop, Christopher M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

ALGORITHMS: ELEMENTS OF STATISTICAL LEARNING

1. Introduction
2. **Overview of supervised learning**
3. **Linear methods for regression**
4. **Linear methods for classification**
5. Basis expansions and regularization
6. **Kernel smoothing methods**
7. Model assessment and selection
8. Model inference and averaging
9. **Additive models, trees, and related methods**
10. **Boosting and additive trees**
11. **Neural networks**
12. **Support vector machines and flexible discriminants**
13. **Prototype methods and nearest-neighbors**
14. **Unsupervised learning**
15. **Random forests**
16. **Ensemble learning**
17. **Undirected graphical models**
18. High-dimensional problems

ALGORITHMS:
PATTERN RECOGNITION AND MACHINE LEARNING

1 Introduction
2 Probability Distributions
3 **Linear Models for Regression**
4 **Linear Models for Classification**
5 **Neural Networks**
6 **Kernel Methods**
7 **Sparse Kernel Machines**
8 **Graphical Models**
9 **Mixture Models and EM**
10 Approximate Inference
11 Sampling Methods
12 Continuous Latent Variables
13 Sequential Data
14 **Combining Models**
Appendix A Data Sets
Appendix B Probability Distributions
Appendix C Properties of Matrices
Appendix D Calculus of Variations
Appendix E Lagrange Multipliers

EXAMPLE OF MATH: SAMPLING

From Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, p526.

11.1. Basic Sampling Algorithms

In this section, we consider some simple strategies for generating random samples from a given distribution. Because the samples will be generated by a computer algorithm they will in fact be *pseudo-random* numbers, that is, they will be deterministically calculated, but must nevertheless pass appropriate tests for randomness. Generating such numbers raises several subtleties (Press *et al.*, 1992) that lie outside the scope of this book. Here we shall assume that an algorithm has been provided that generates pseudo-random numbers distributed uniformly over $(0, 1)$, and indeed most software environments have such a facility built in.

11.1.1 Standard distributions

We first consider how to generate random numbers from simple nonuniform distributions, assuming that we already have available a source of uniformly distributed random numbers. Suppose that z is uniformly distributed over the interval $(0, 1)$, and that we transform the values of z using some function $f(\cdot)$ so that $y = f(z)$. The distribution of y will be governed by

$$p(y) = p(z) \left| \frac{dz}{dy} \right| \tag{11.5}$$

where, in this case, $p(z) = 1$. Our goal is to choose the function $f(z)$ such that the resulting values of y have some specific desired distribution $p(y)$. Integrating (11.5) we obtain

$$z = h(y) \equiv \int_{-\infty}^{y} p(\widehat{y}) \, \mathrm{d}\widehat{y} \tag{11.6}$$

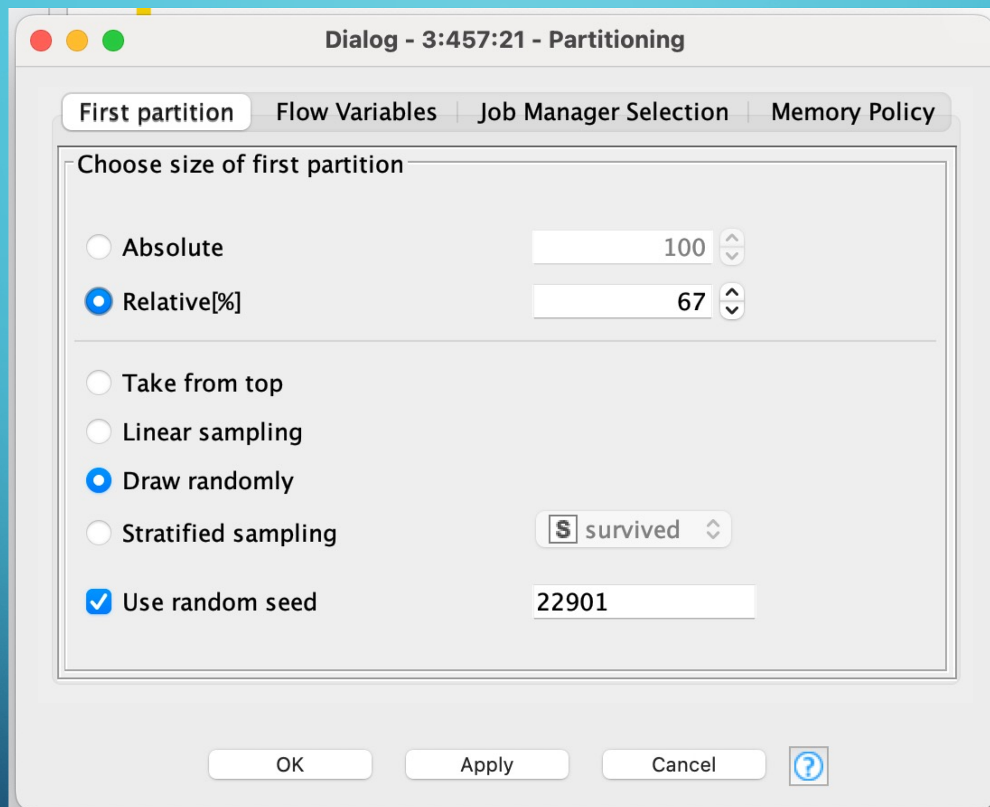
Exercise 11.2

which is the indefinite integral of $p(y)$. Thus, $y = h^{-1}(z)$, and so we have to transform the uniformly distributed random numbers using a function which is the inverse of the indefinite integral of the desired distribution. This is illustrated in Figure 11.2.

Consider for example the *exponential distribution*

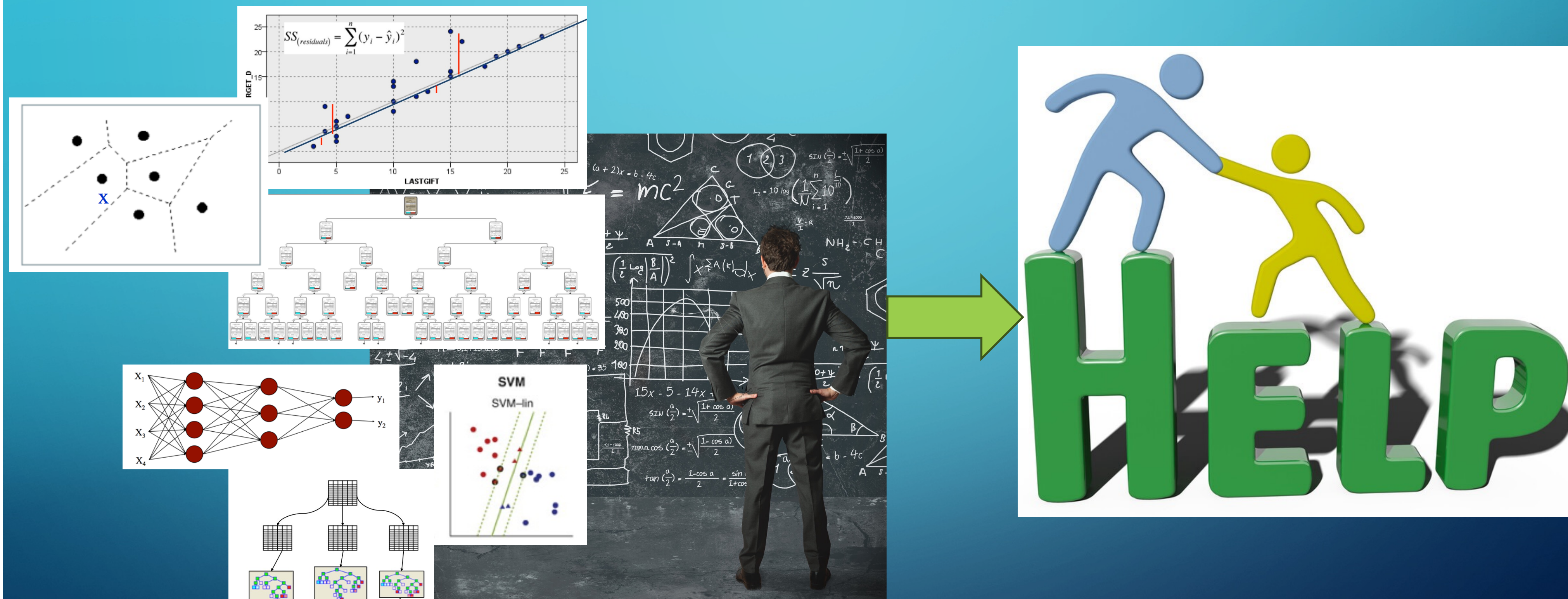
$$p(y) = \lambda \exp(-\lambda y) \tag{11.7}$$

WHAT WE ACTUALLY DO: SAMPLING



PURPOSE OF THE 12 IDEAS

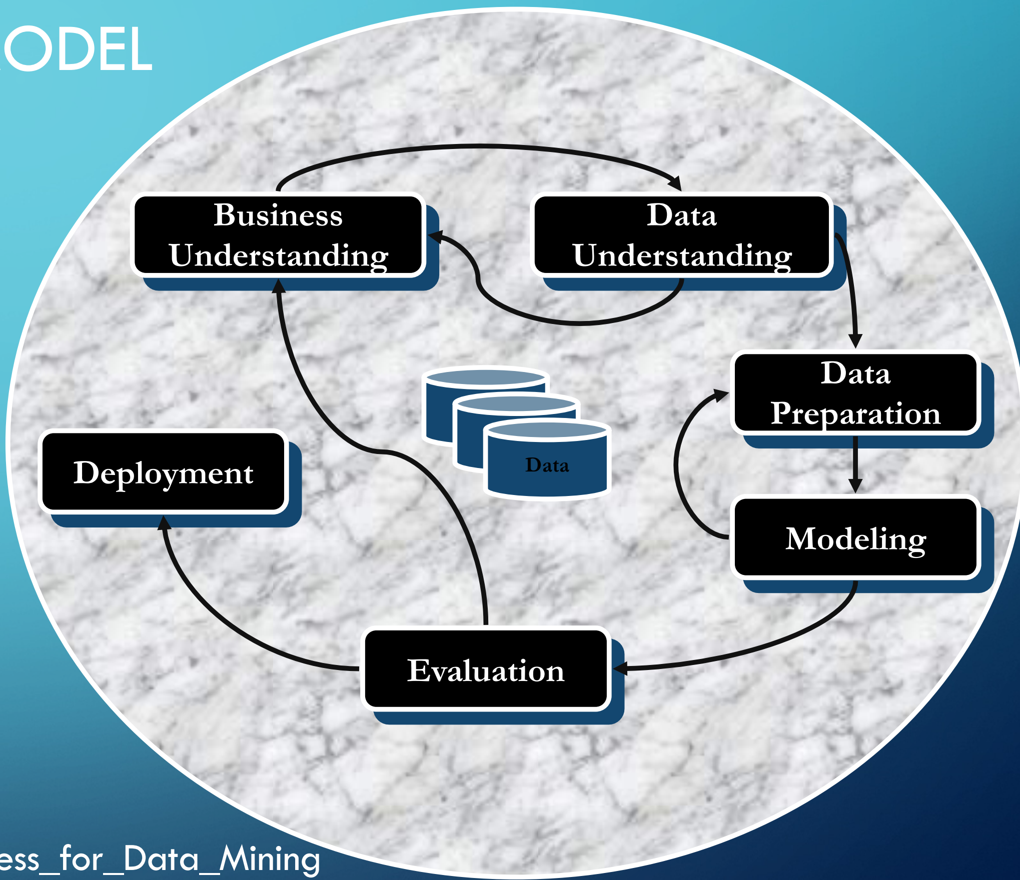
1. CLARIFY THE CONFUSING



2. PROVIDE INSIGHTS THAT MIGHT BE MISSED

WHAT DO DATA SCIENTISTS DO?
THE "CRISP-DM" PROCESS MODEL

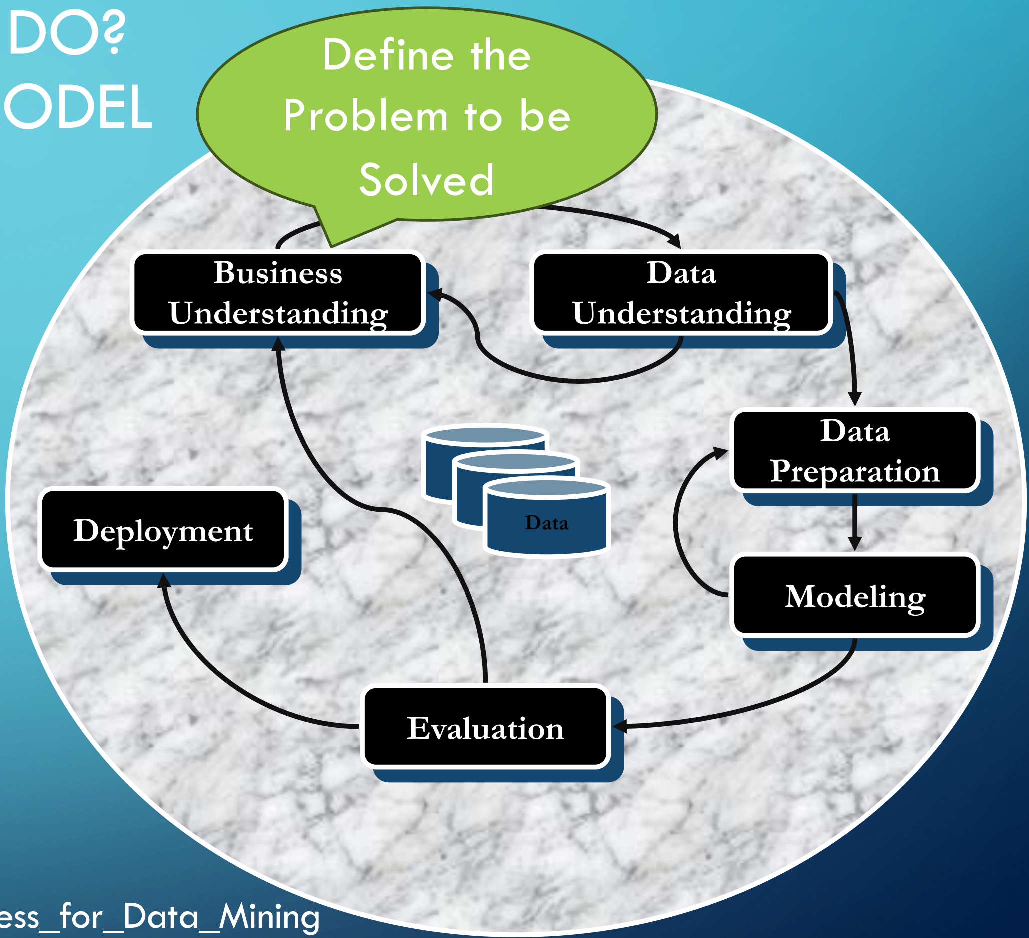
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

WHAT DO DATA SCIENTISTS DO? THE "CRISP-DM" PROCESS MODEL

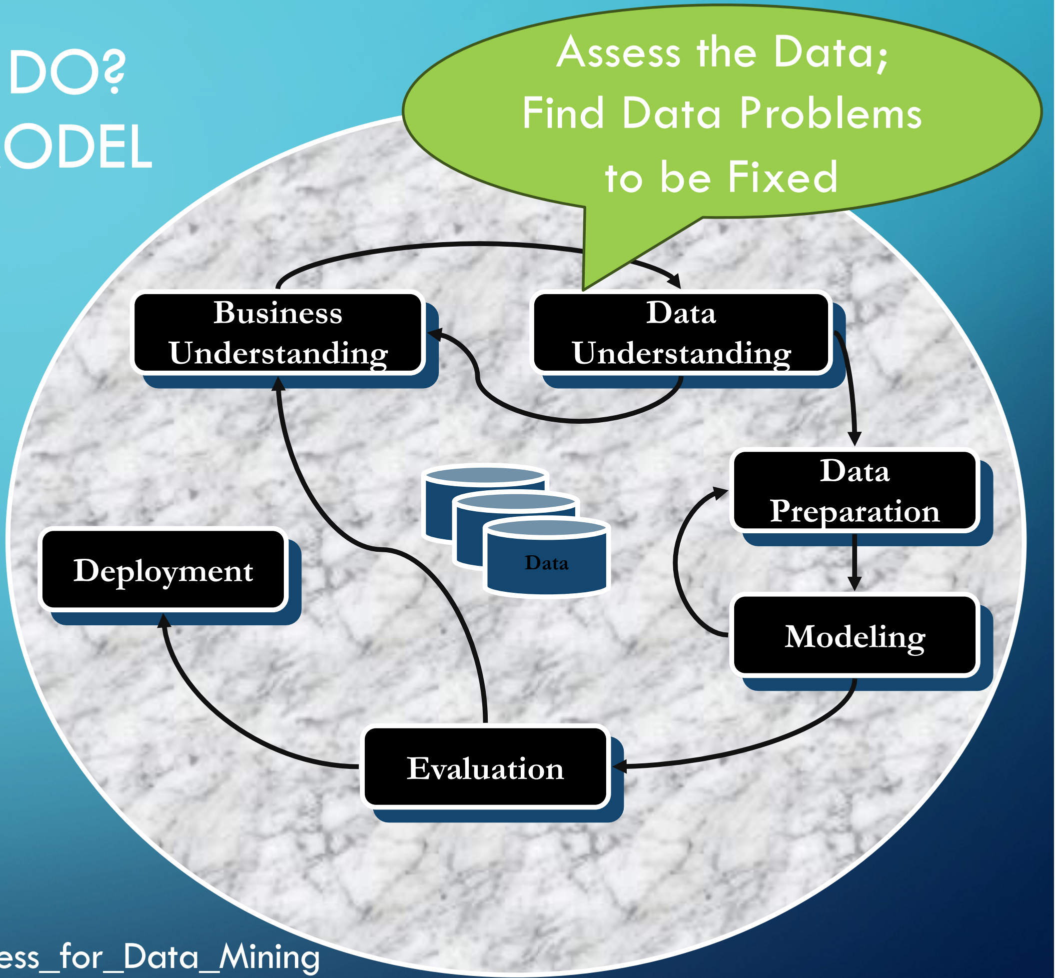
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

WHAT DO DATA SCIENTISTS DO? THE "CRISP-DM" PROCESS MODEL

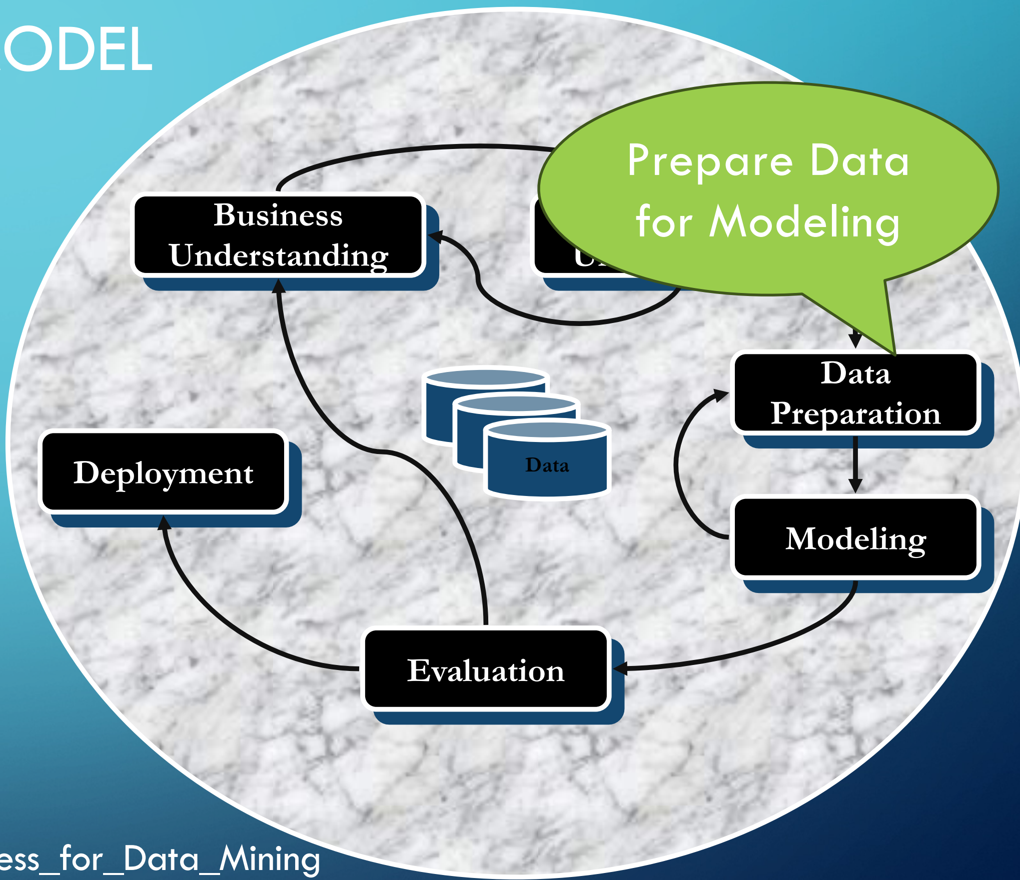
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

WHAT DO DATA SCIENTISTS DO? THE “CRISP-DM” PROCESS MODEL

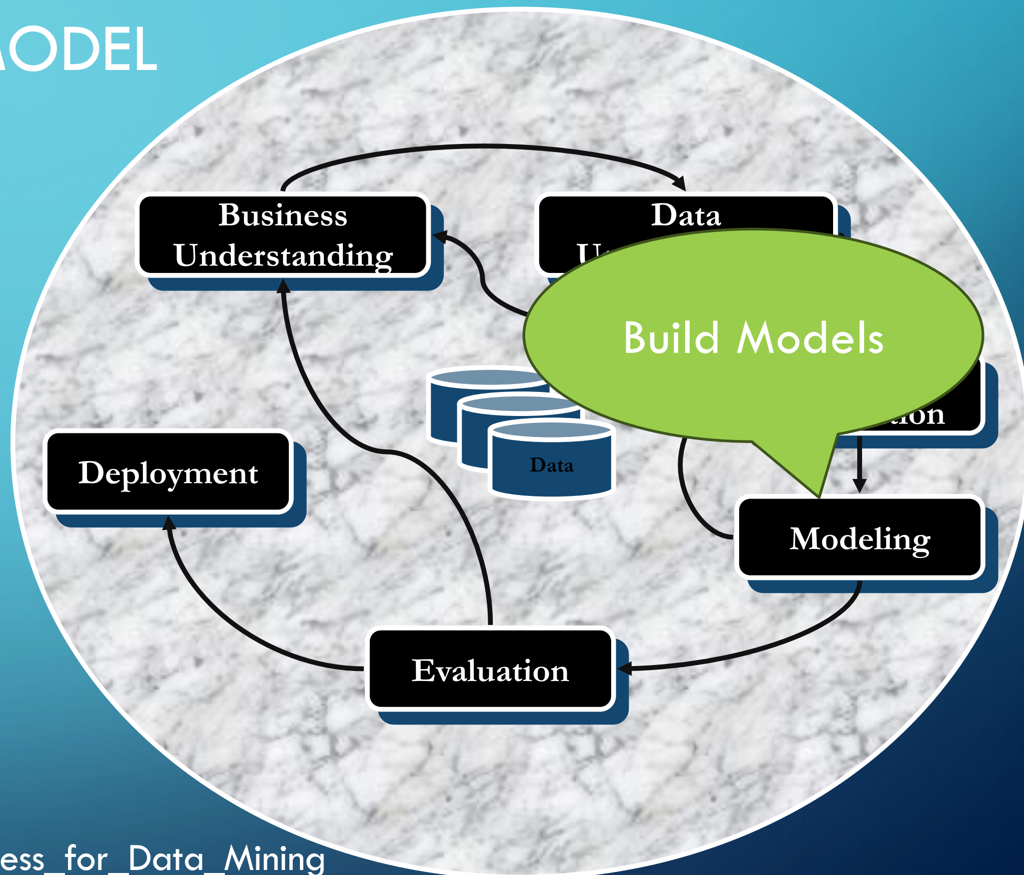
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager’s Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

WHAT DO DATA SCIENTISTS DO? THE "CRISP-DM" PROCESS MODEL

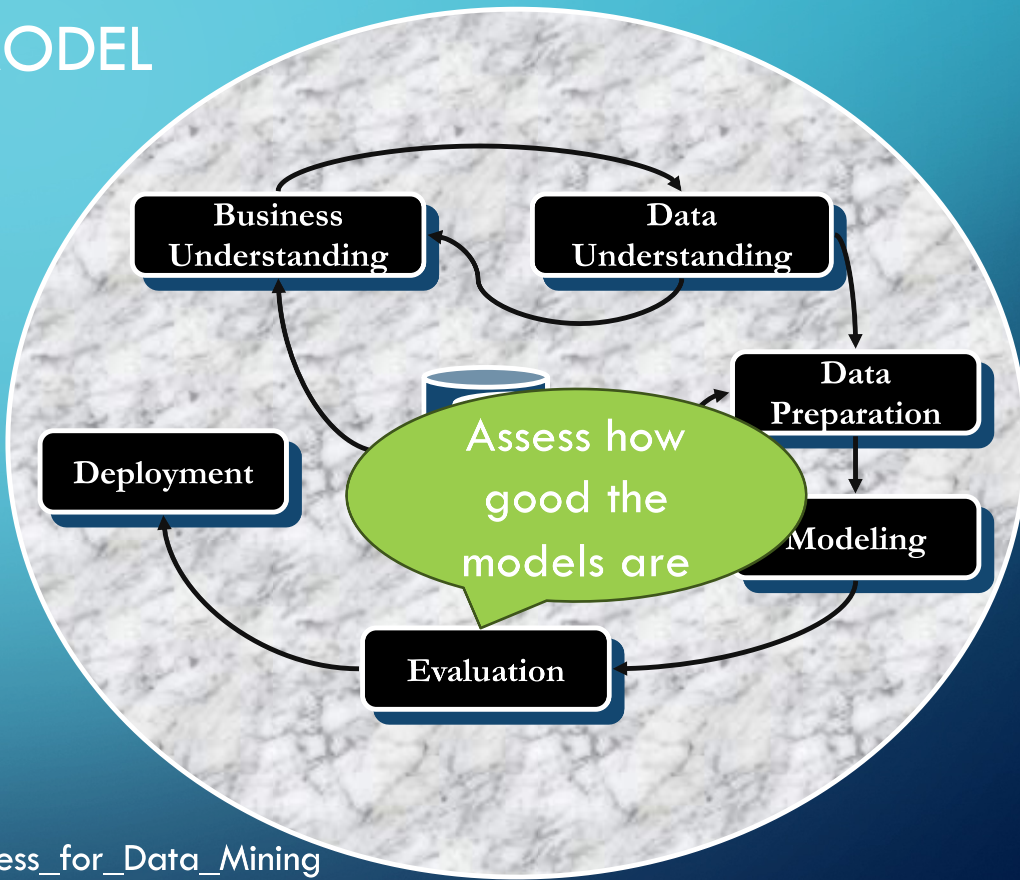
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

WHAT DO DATA SCIENTISTS DO? THE "CRISP-DM" PROCESS MODEL

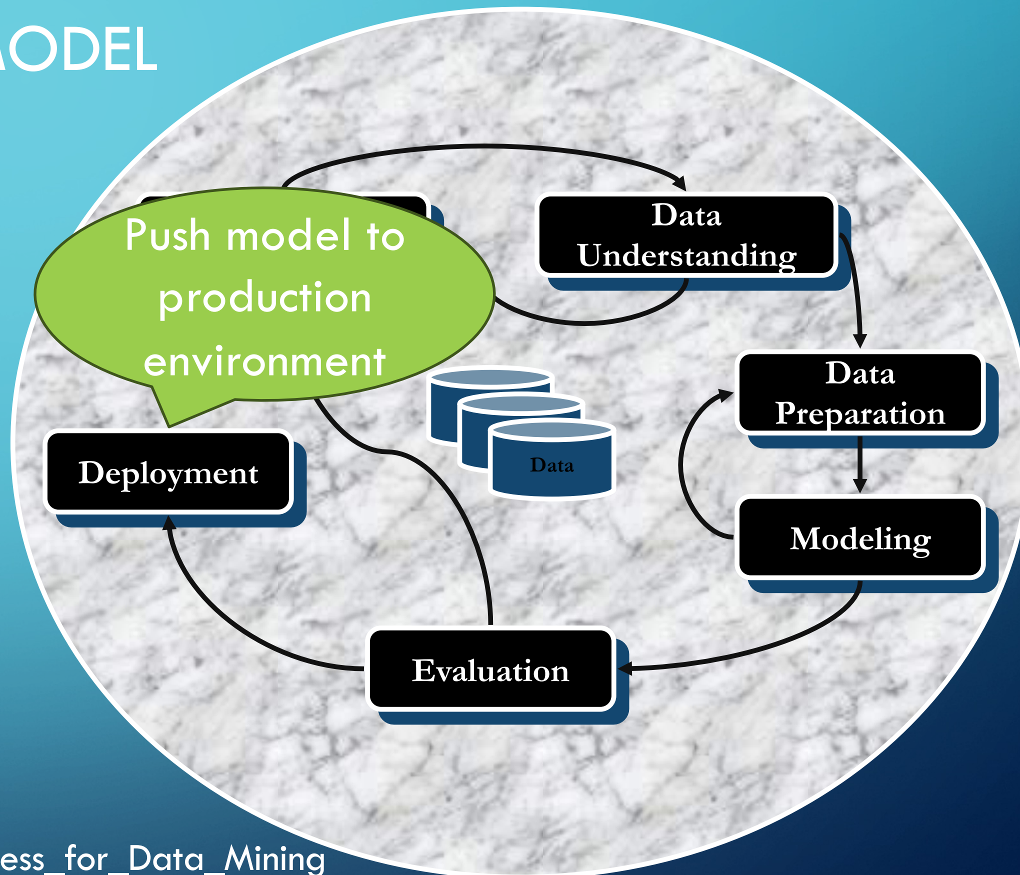
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

WHAT DO DATA SCIENTISTS DO? THE "CRISP-DM" PROCESS MODEL

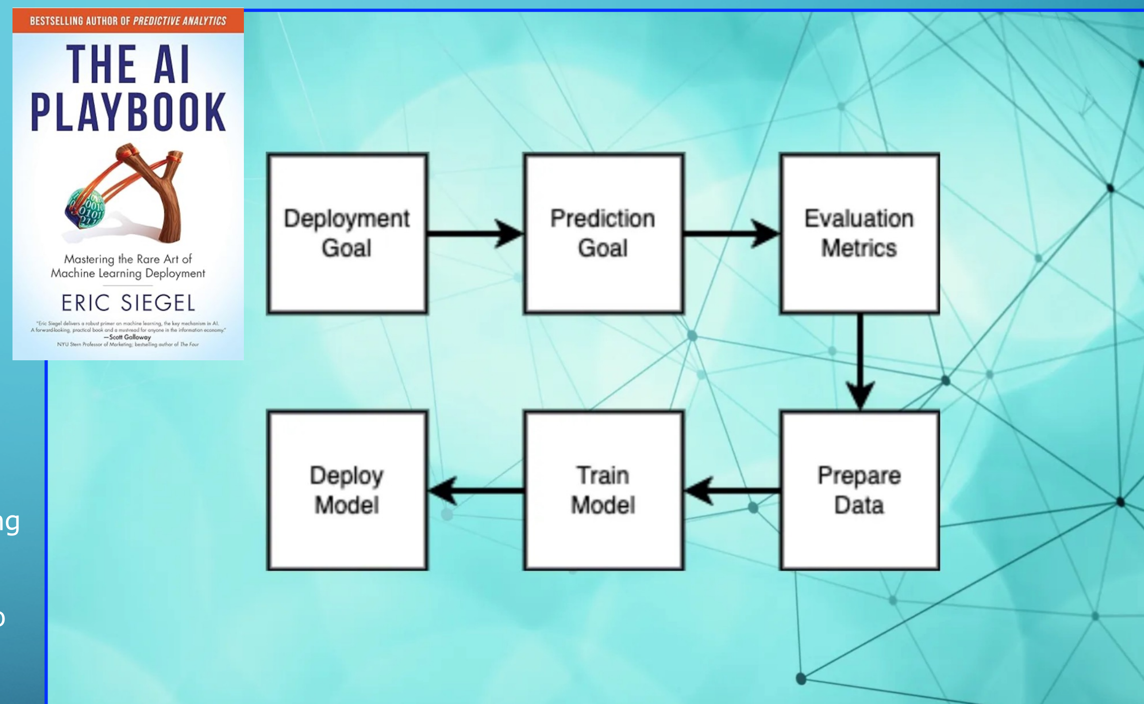
- CRoss-Industry Standard Process Model for Data Mining – CRISP-DM
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

MODERNIZED VERSIONS OF CRISP-DM: BIZML FRAMEWORK

- 1- **Establish the deployment goal**: Define clearly how ML will affect your operations
- 2- **Establish the prediction goal**: Determine what the model will predict and how it relates to the deployment goal
- 3- **Establish the evaluation metrics**: Determine the metrics that matter and the performance level required to achieve the deployment goal
- 4- **Prepare the data**: Define what the data must look like and prepare the datasets
- 5- **Train the model**: Use the data to train your machine learning model
- 6- **Deploy the model**: Integrate the model into your product to make predictions on new data coming from business operations



https://bdtechtalks.com/2024/02/05/ai-playbook-bizml-review/

TWELVE IDEAS THAT TRANSFORM OUR MODELS: WHAT DATA SCIENTISTS DISCOVER THEY ACTUALLY DO

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models

Modeling

6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data

Evaluation

9. Be Prepared to Quantify Machine Learning Model Prediction Stability.
10. Quantify Variable Importance Related to Its Use When Deployed

Deployment

11. Prepare for Model Obsolescence: When and How to Rebuild Models

TWELVE IDEAS THAT TRANSFORM OUR MODELS

Business Understanding

Data Preparation

Modeling

Evaluation

Deployment

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

TWELVE IDEAS THAT TRANSFORM OUR MODELS

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models

Modeling

Evaluation

Deployment

TWELVE IDEAS THAT TRANSFORM OUR MODELS

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models

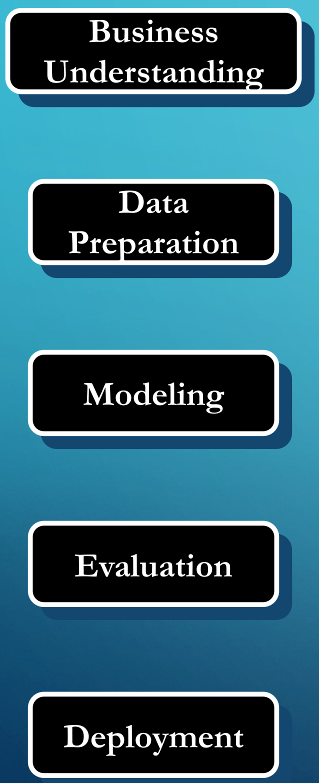
Modeling

6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data

Evaluation

Deployment

TWELVE IDEAS THAT TRANSFORM OUR MODELS



1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.
3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models
6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data
9. Be Prepared to Quantify Machine Learning Model Prediction Stability.
10. Quantify Variable Importance Related to Its Use When Deployed

TWELVE IDEAS THAT TRANSFORM OUR MODELS

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models

Modeling

6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data

Evaluation

9. Be Prepared to Quantify Machine Learning Model Prediction Stability.
10. Quantify Variable Importance Related to Its Use When Deployed

Deployment

11. Prepare for Model Obsolescence: When and How to Rebuild Models

WHAT DATA SCIENTISTS DISCOVER THEY ACTUALLY DO

Business Understanding	1. Define the Target Variable Carefully and Correctly 2. Connect the measure of model success to the model's Business Objective.
Data Preparation	3. Account for Dummy Variables appropriately (differently from Other Numerical Variables) 4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy 5. Beware of the Destructive Influence of Irrelevant Variables In Models
Modeling	6. Remember that Even Robust Ensembles Can Overfit 7. Account for High Class Imbalance with the Algorithm in Mind 8. Beware of Ways Even Advanced Algorithms are Fooled By Data
Evaluation	9. Be Prepared to Quantify Machine Learning Model Prediction Stability. 10. Quantify Variable Importance Related to Its Use When Deployed
Deployment	11. Prepare for Model Obsolescence: When and How to Rebuild Models

CRISP-DM:
BUSINESS UNDERSTANDING STEPS

Define the Business Objectives for the analytics projects in the language of the business

Translate the business objectives into the analytics objectives

EXAMPLE 1: INVOICE "FRAUD"

- **Business Objective:** provide invoices for investigators to examine that are "suspicious"
- **Machine Learning Objective:**
  - Create data: 1 record per invoice
  - Create label for each invoice: "Suspicious" vs. "Not-Suspicious" (1/0 Target variable)

EXAMPLE 2: TAX RETURN NON-COMPLIANCE

- **Business Objective:** provide auditors good workload for audit—tax returns that are likely to have non-compliance.
- **Target Variable:** "Compliant" vs. "Non-Compliant" (1/0 Target variable)
  - But, what is "Compliant"?

EXAMPLE 3: NON-PROFIT LAPSED DONOR RECOVERY

- **Business Objective:** identify which lapsed donors can be recovered
- **Target Variable:** "Recoverable vs. Not Recoverable" (1/0 Target variable) and / or $$ value of recovery

Paralyzed Veterans of America

https://kdd.org/kdd-cup/view/kdd-cup-1998/Data

WHAT DATA SCIENTISTS DISCOVER THEY ACTUALLY DO

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models

Modeling

6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data

Evaluation

9. Be Prepared to Quantify Machine Learning Model Prediction Stability.
10. Quantify Variable Importance Related to Its Use When Deployed

Deployment

11. Prepare for Model Obsolescence: When and How to Rebuild Models

CRISP-DM:
BUSINESS UNDERSTANDING STEPS

- Define what a good outcome is in the language of the business

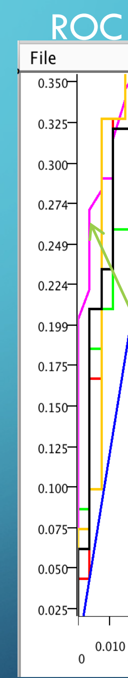
- Translate the business success criterion into an analytics criterion that matches as closely as possible

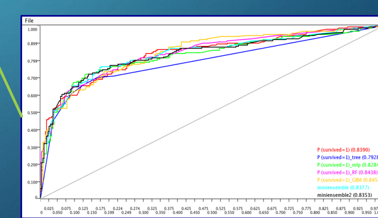
EXAMPLE 1: INVOICE "FRAUD" SUCCESS

- **Error Metric for Model Selection and Estimate of Model Effectiveness:** Classification Accuracy

EXAMPLE 1: MITIGATING CIRCUMSTANCES

- **But**: The investigators could only process 100 invoices per month
- **Operational Implication: Only deliver top 100 scoring invoices per month that have high chance of being suspicious**
- **Model Success Implication:** Scores for the 1M+ invoices after the top 100 are **irrelevant.**
  - Winning algorithm will be the one that identifies the extreme tail of the distribution the best.





EXAMPLE 2: TAX RETURN "NON-COMPLIANCE" SUCCESS

- **Error Metric for Model Selection and Estimate of Model Effectiveness:** Classification Accuracy

EXAMPLE 2: MITIGATING CIRCUMSTANCES

- **But**: For the IRS, most important outcome is taxpaying entities who owe IRS
  - Avoid "no-change"
  - Avoid small changes where ROI is small or even negative
- **Operational Implication:** Workflow Tiers
  - **Model outcome 1: send returns to field with positive expected tax change > $N**
  - **Model Outcome 2: if more returns needed, send returns with negative expected tax change > $M**
  - **All other returns NOT bsent to field**

EXAMPLE 3: LAPSED DONOR RECOVERY SUCCESS

- **Error Metric for Model Selection and Estimate of Model Effectiveness:** ROI

EXAMPLE 3 : MITIGATING CIRCUMSTANCES

- **But**: For the PVA, contacts (treatment) has a cost
  - $N per contact means the model should identify those value exceeds $N
- **Operational Implication:** ROI
  - **Pick model that maximizes the cumulative net revenue of a set of lapsed donors**
  - **No machine learning algorithm does this: we have to impose this metric after the models minimize/maximize"error" according their algorithm**

WHAT DATA SCIENTISTS DISCOVER THEY ACTUALLY DO

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. **Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy**
5. Beware of the Destructive Influence of Irrelevant Variables In Models

Modeling

6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data

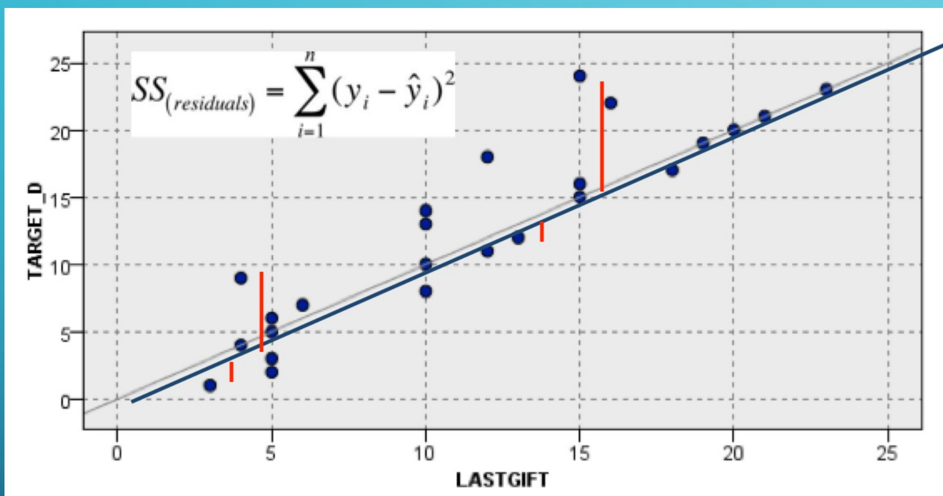
Evaluation

9. Be Prepared to Quantify Machine Learning Model Prediction Stability.
10. Quantify Variable Importance Related to Its Use When Deployed

Deployment

11. Prepare for Model Obsolescence: When and How to Rebuild Models

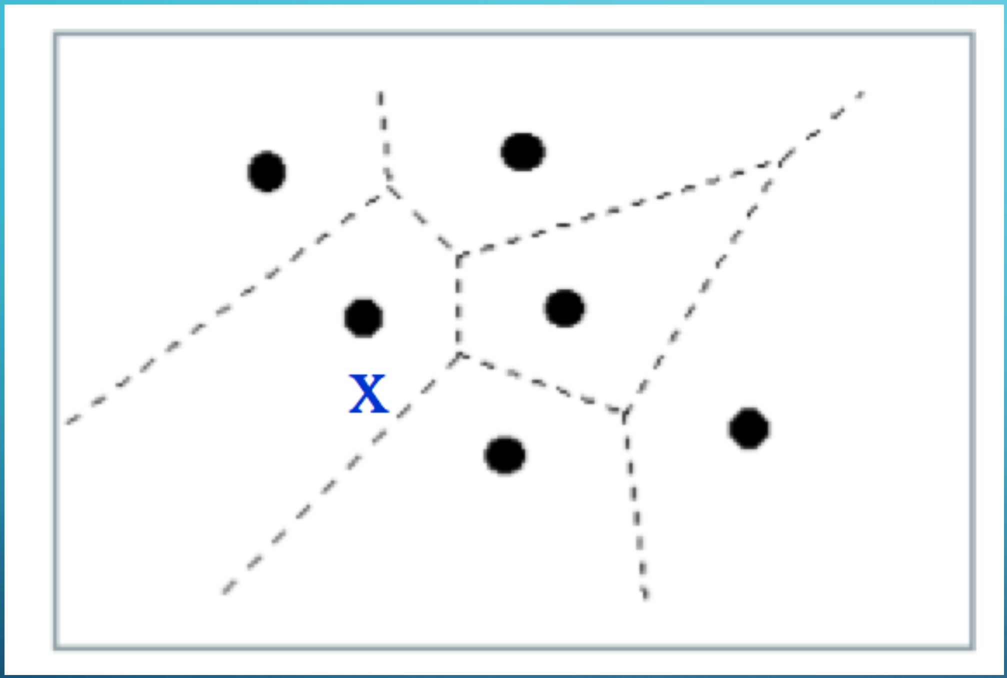
DOES NOT NATURALLY FIND INTERACTIONS: LINEAR / LOGISTIC REGRESSION



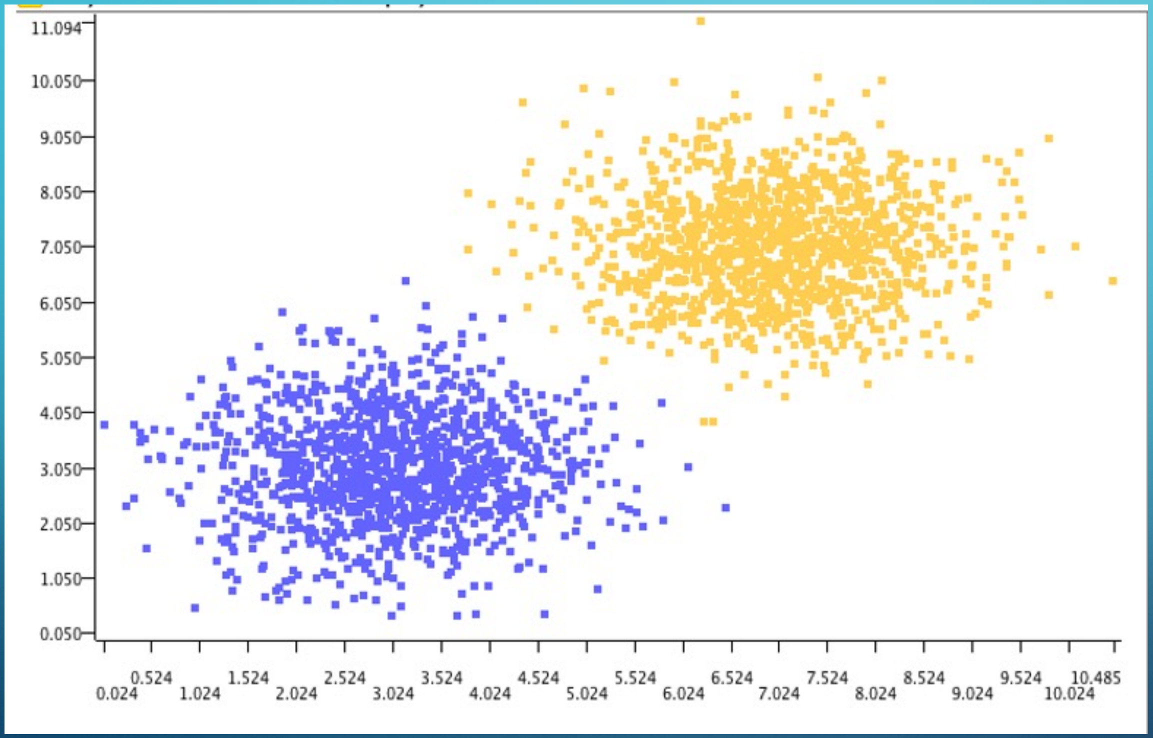
$$y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_mx_m$$

output | bias term | linear combination of inputs (w are weights)

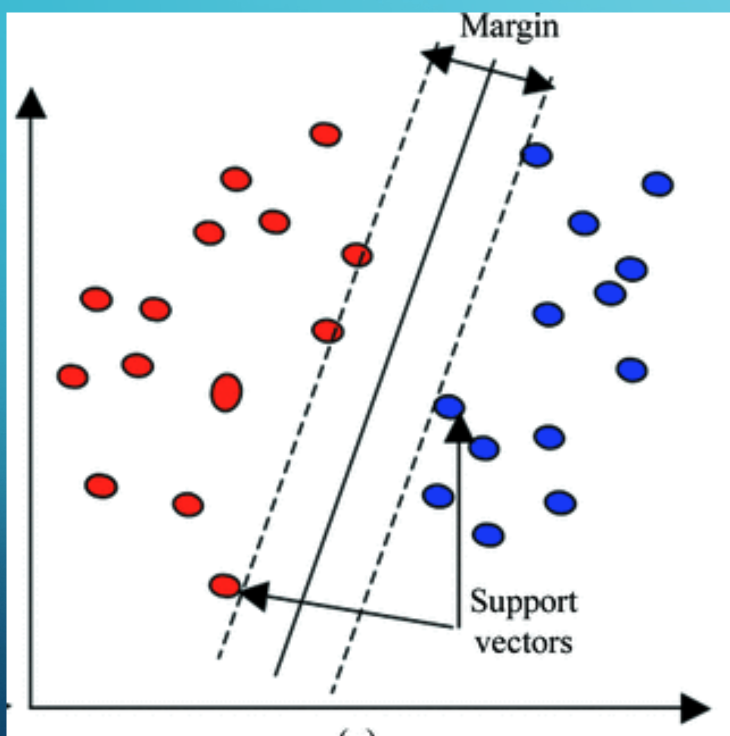
DOES NOT NATURALLY FIND INTERACTIONS: K-NEAREST NEIGHBOR

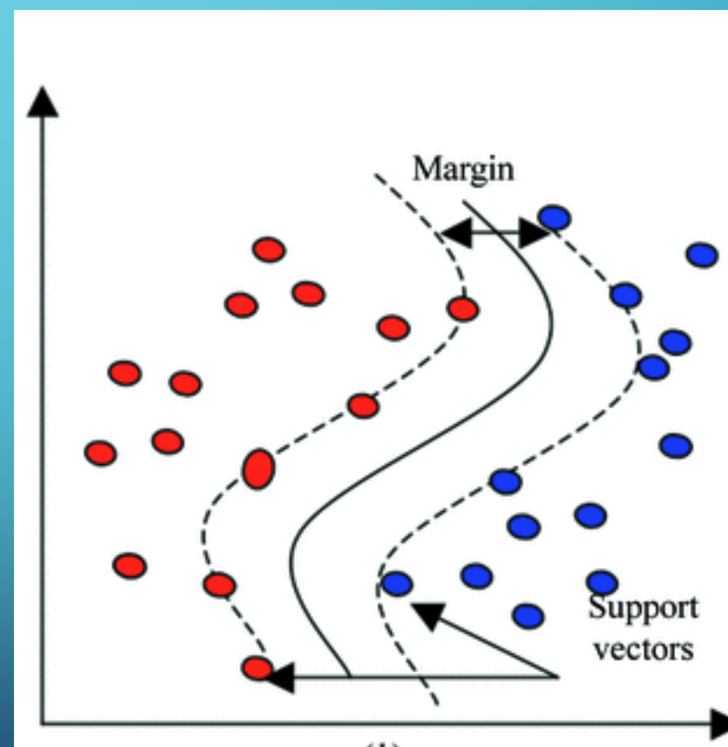


DOES NOT NATURALLY FIND INTERACTIONS: K MEANS CLUSTERING

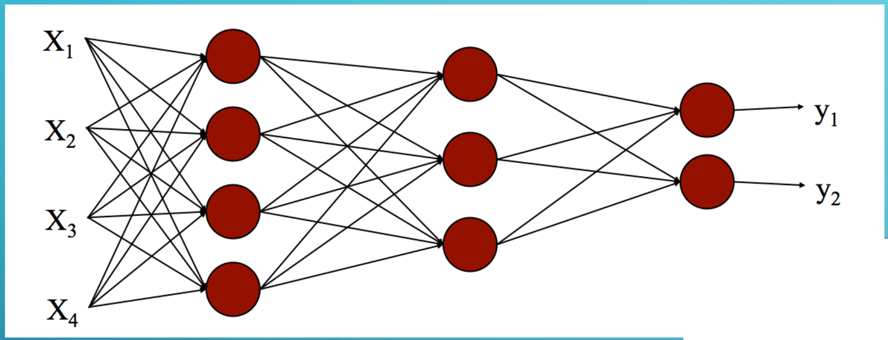


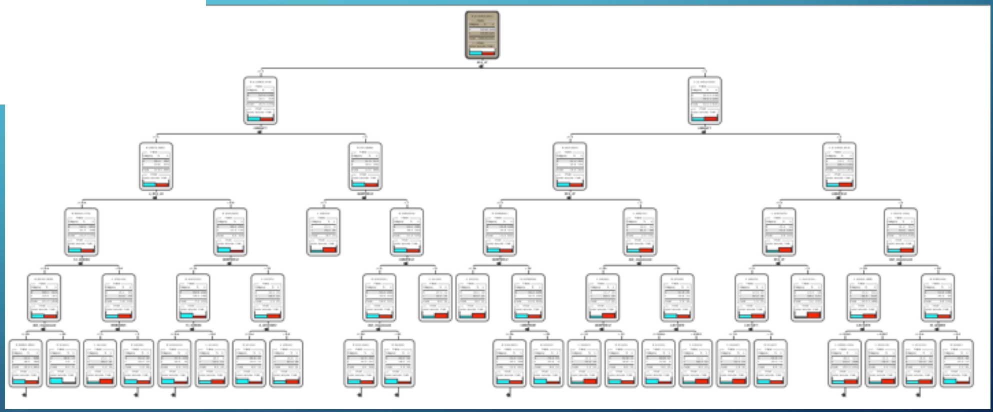
DOES NOT NATURALLY FIND INTERACTIONS: SUPPORT VECTOR MACHINES





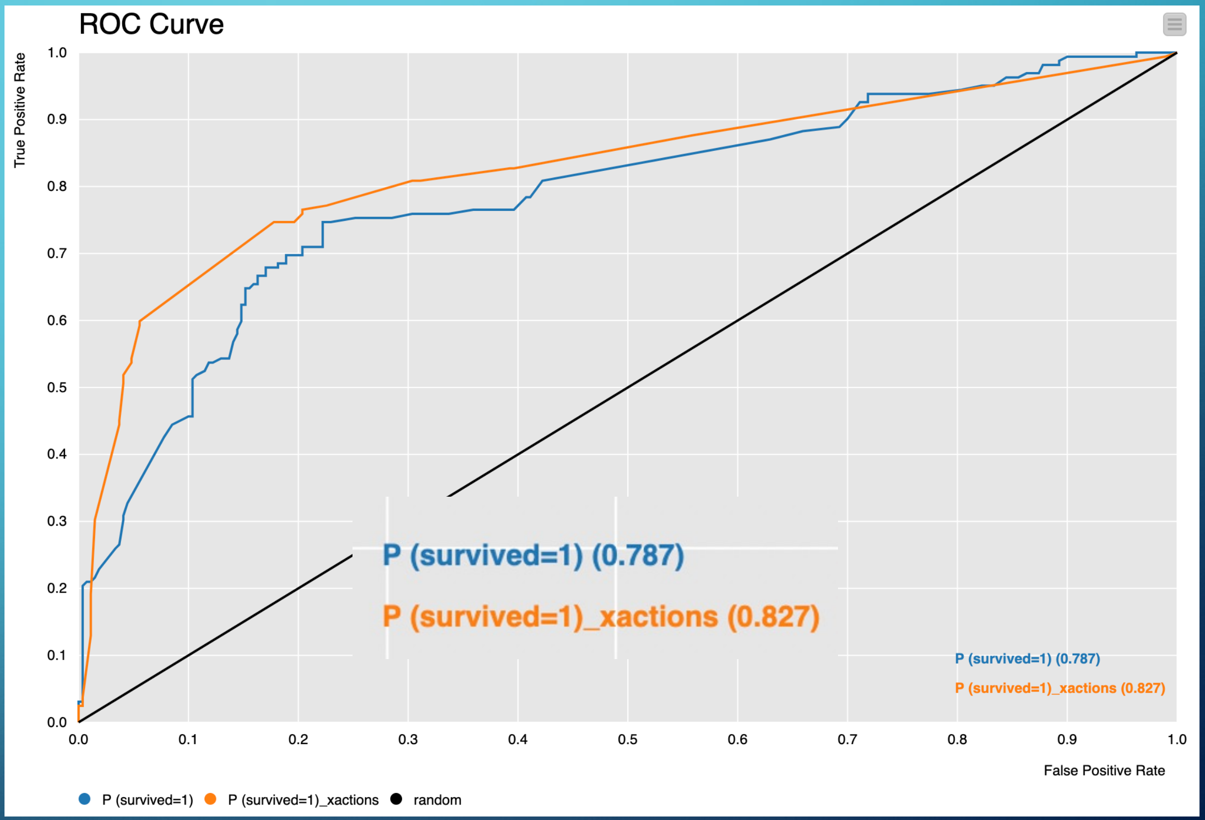
DOES NATURALLY FIND INTERACTIONS: SUPPORT VECTOR MACHINES





INTERACTIONS MATTER: LOGISTIC REGRESSION

Columns: 20	Column Type
survived	String
age	Number (double)
pclass_1	Number (integer)
pclass_2	Number (integer)
male	Number (integer)
embarked_C	Number (integer)
embarked_S	Number (integer)
cabin_B	Number (integer)
cabin_E	Number (integer)
cabin_D	Number (integer)
cabin_A	Number (integer)
cabin_U	Number (integer)
cabin_F	Number (integer)
sibsp_0	Number (integer)
sibsp_2_3	Number (integer)
sibsp_4+	Number (integer)
parch_0	Number (integer)
parch_2_3	Number (integer)
parch_4+	Number (integer)
fare_log10	Number (double)

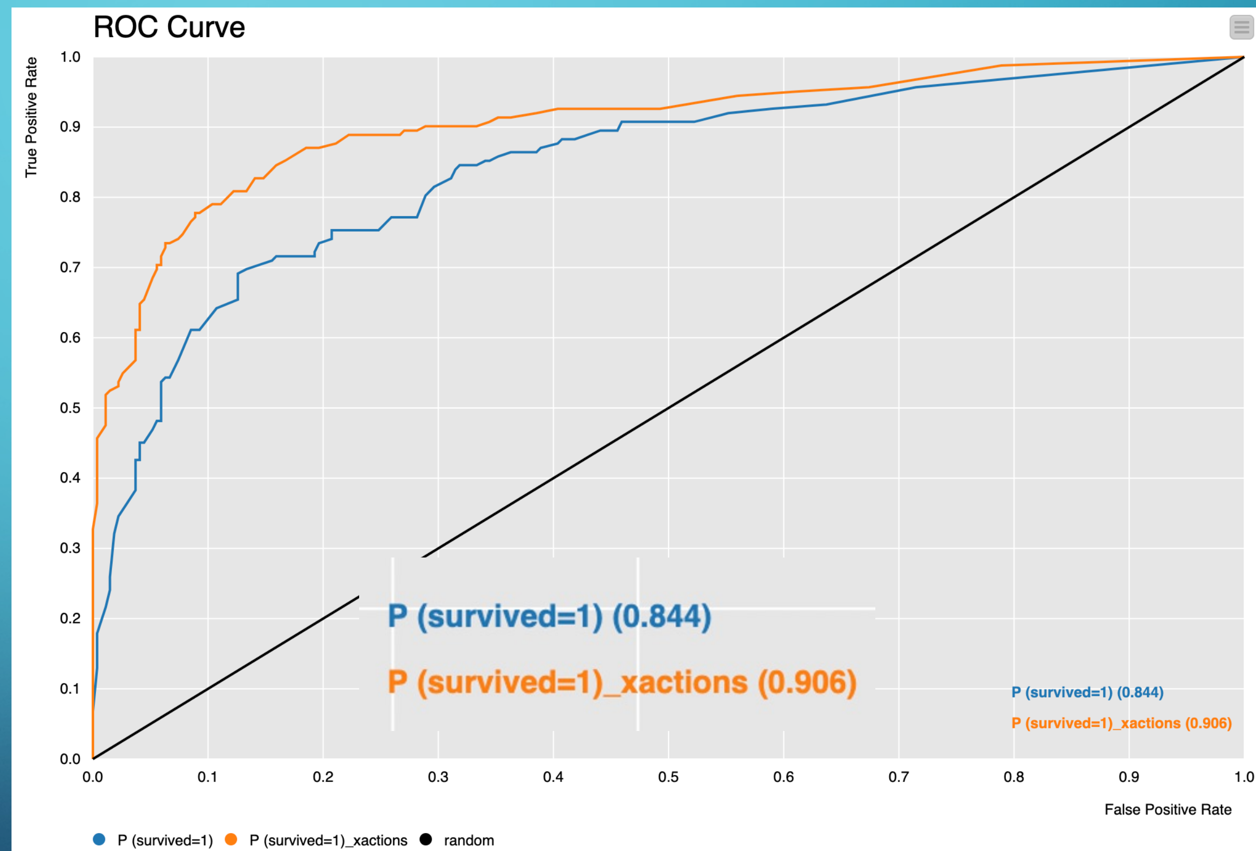


Add Feature:
Pclass_2 and female

INTERACTIONS MATTER: EVEN RANDOM FORESTS!

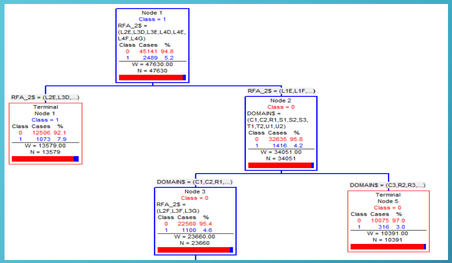
Columns: 20	Column Type
survived	String
age	Number (double)
pclass_1	Number (integer)
pclass_2	Number (integer)
male	Number (integer)
embarked_C	Number (integer)
embarked_S	Number (integer)
cabin_B	Number (integer)
cabin_E	Number (integer)
cabin_D	Number (integer)
cabin_A	Number (integer)
cabin_U	Number (integer)
cabin_F	Number (integer)
sibsp_0	Number (integer)
sibsp_2_3	Number (integer)
sibsp_4+	Number (integer)
parch_0	Number (integer)
parch_2_3	Number (integer)
parch_4+	Number (integer)
fare_log10	Number (double)

Add Feature:
Pclass_2 and
female



INTERACTION DETECTION

- Trees: build 2-level trees
- Use the linear/logistic regression algorithm itself, loop over all 2-way interactions
- Association rules: build 2-antecedent rules



Statistics on Logistic Regression

Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
1	NGIFTALL	0.0239	0.0034	7.0132	2.33E-12
	LASTGIFT	-0.0093	0.0028	-3.3185	0.0009
	Constant	-0.0985	0.0694	-1.419	0.1559

Log-likelihood = -3,322.7813
Number of iterations = 8

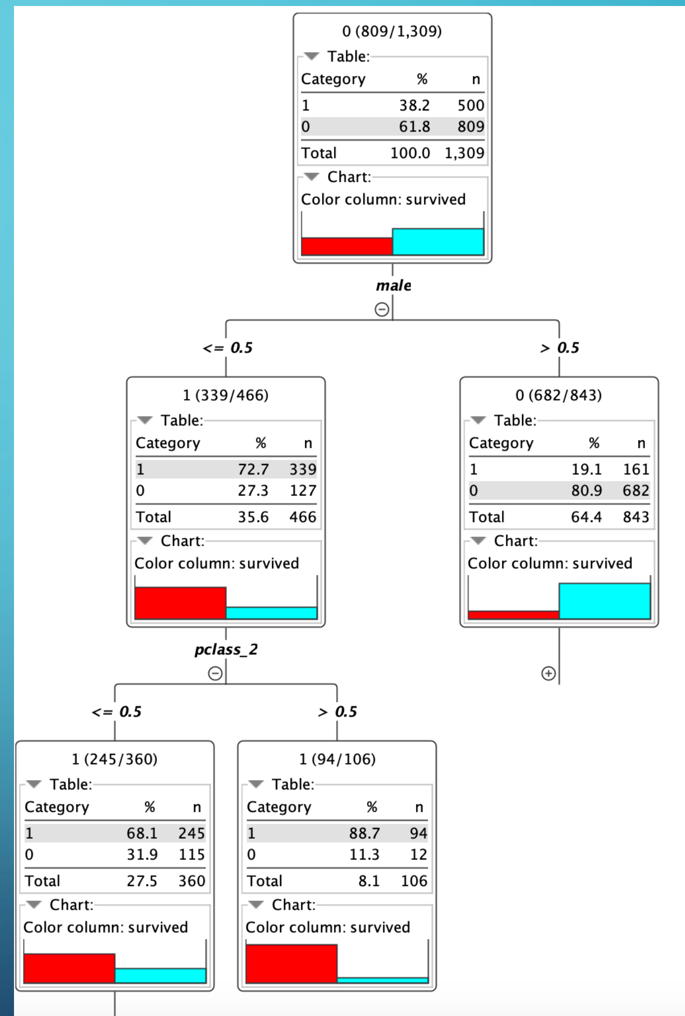
A	B	C	D	E	F	G
row ID	Support	Confidence	Lift	Consequent	Split Value 1	Split Value 2
rule0	0.01004613	34.1176471	3.47151569	DaysToNextPurchase_le_60	AssetCount_51-100	PurchaseFlag_eq_true
rule8	0.01004613	5.30164534	4.8576988	DaysToNextPurchase_4-7	ChannelEngagement_8000-20000	DaysSinceLastPurchase_31-60
rule18	0.01004613	3.17982456	1.55197428	DaysToNextPurchase_le_7	AverageDaysBetweenVisits_31-60	DaysSinceLastPurchase_null
rule22	0.01004613	3.17982456	1.55197428	DaysToNextPurchase_le_7	PriorPurchase_eq_false	AverageDaysBetweenVisits_31-60
rule26	0.01004613	8.49194729	4.14465751	DaysToNextPurchase_le_7	ChannelEngagement_8000-20000	DaysSinceLastPurchase_ge_91
rule30	0.01004613	16.3380282	4.47697953	DaysToNextPurchase_le_14	AssetCount_51-100	DaysSinceLastPurchase_null
rule34	0.01004613	16.3380282	4.47697953	DaysToNextPurchase_le_14	PriorPurchase_eq_false	AssetCount_51-100
rule38	0.01004613	3.9862543	1.09232147	DaysToNextPurchase_le_14	DaysSinceLastPurchase_ge_91	DaysSinceLastVisit_61-90
rule42	0.01004613	7.03030303	1.0945618	DaysToNextPurchase_le_30	VisitQuality_1-1000	AverageDaysBetweenVisits_le_1
rule46	0.01004613	13.4570766	2.09515889	DaysToNextPurchase_le_30	AssetCount_11-20	DaysSinceLastVisit_8-14
rule50	0.01004613	9.78077572	0.99520686	DaysToNextPurchase_le_60	ChannelEngagement_1000-3000	AssetCount_6-10

FINDING INTERACTIONS USING DECISION TREES

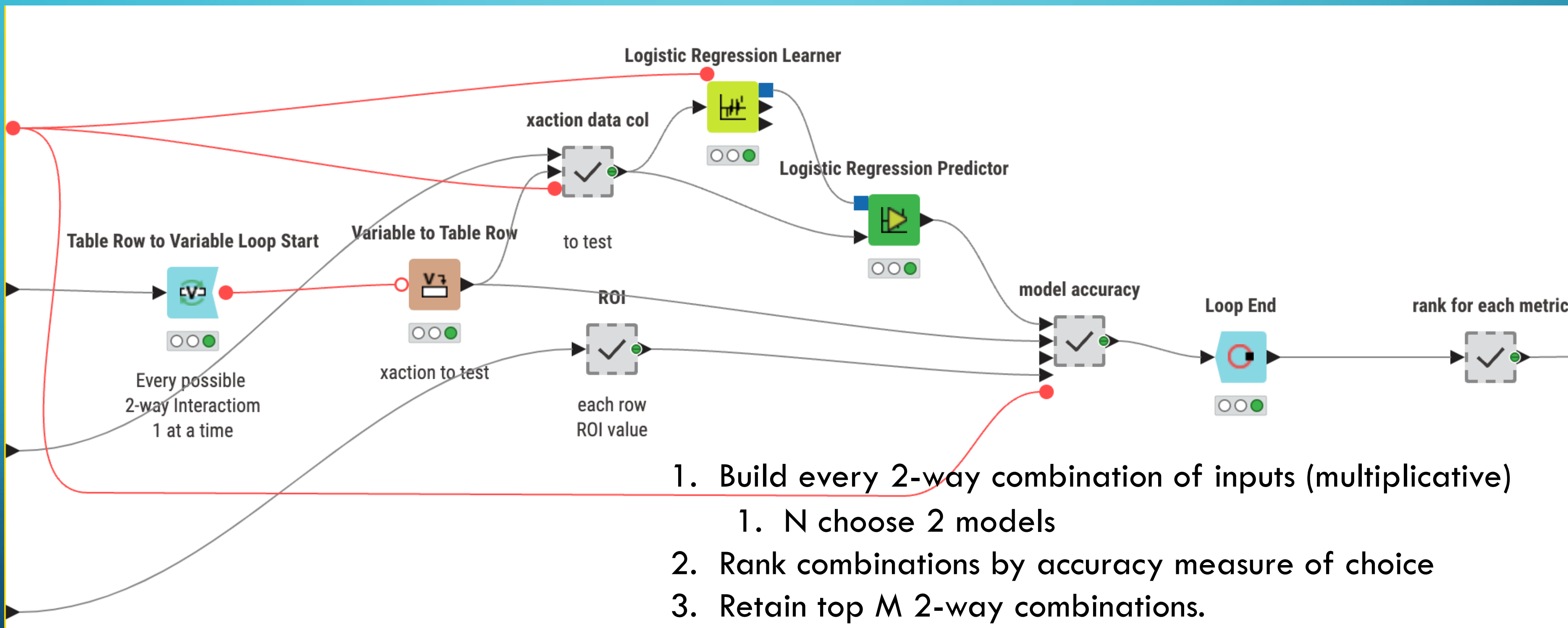
Columns: 20	Column Type
survived	String
age	Number (double)
pclass_1	Number (integer)
pclass_2	Number (integer)
male	Number (integer)
embarked_C	Number (integer)
embarked_S	Number (integer)
cabin_B	Number (integer)
cabin_E	Number (integer)
cabin_D	Number (integer)
cabin_A	Number (integer)
cabin_U	Number (integer)
cabin_F	Number (integer)
sibsp_0	Number (integer)
sibsp_2_3	Number (integer)
sibsp_4+	Number (integer)
parch_0	Number (integer)
parch_2_3	Number (integer)
parch_4+	Number (integer)
fare_log10	Number (double)

If $pclass_2$ = 1
AND $female$ = 1

then survived (87.2%)



FINDING INTERACTIONS VIA EXHAUSTIVE SEARCH: FACTORIAL DESIGN

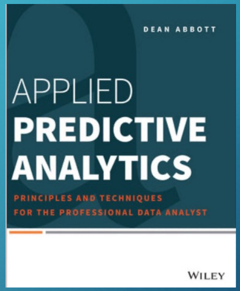


1. Build every 2-way combination of inputs (multiplicative)
    1. N choose 2 models
2. Rank combinations by accuracy measure of choice
3. Retain top M 2-way combinations.

THE PROBLEM WITH FACTORIAL DESIGN

Table 4-16: Number of Two-Way Interaction Combinations

NUMBER OF VARIABLES	NUMBER OF POSSIBLE TWO-WAY INTERACTIONS
5	10
10	45
50	1,225
100	4,950
500	124,750
1000	499,500

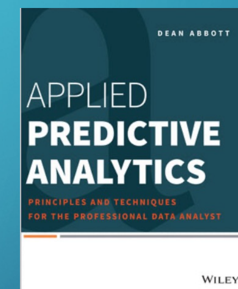


https://www.wiley.com/en-us/Applied+Predictive+Analytics%3A+Principles+and+Techniques+for+the+Professional+Data+Analyst-p-9781118727966

THE PROBLEM WITH FACTORIAL DESIGN

Table 4-16: Number of Two-Way Interaction Combinations

NUMBER OF VARIABLES	NUMBER OF POSSIBLE TWO-WAY INTERACTIONS
5	10
10	45
50	1,225
100	4,950
500	124,750
1000	499,500

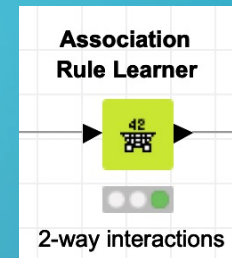


https://www.wiley.com/en-us/Applied+Predictive+Analytics%3A+Principles+and+Techniques+for+the+Professional+Data+Analyst-p-9781118727966

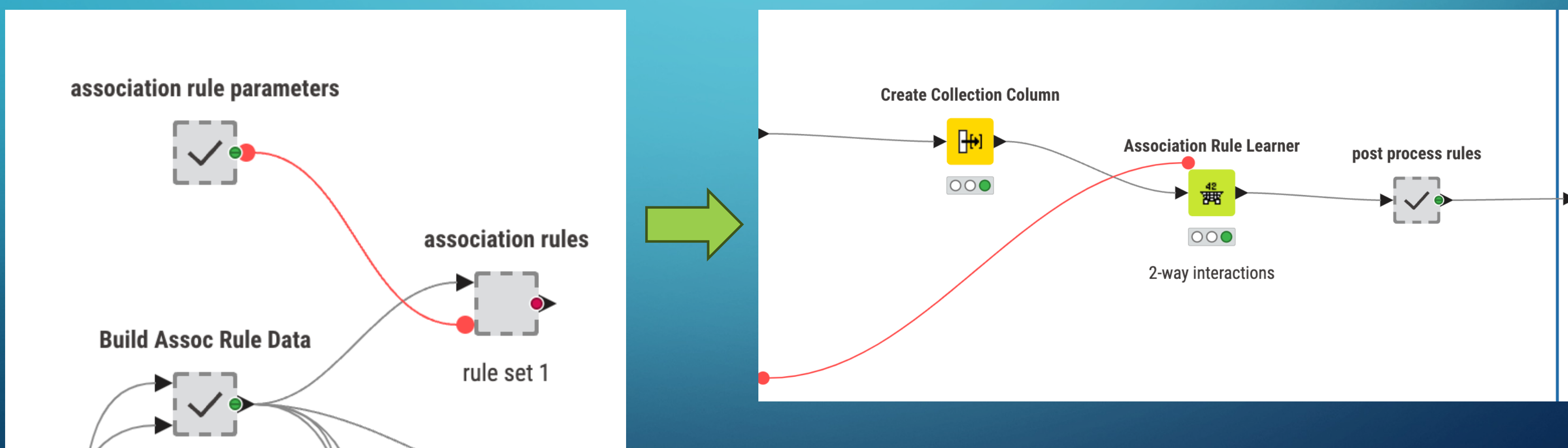
Example:

Algorithm	Records	Inputs*	Combos	Time
Logistic	95,412	60	1,770	5 hours

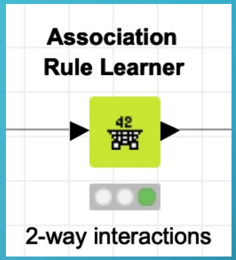
USE ASSOCIATION RULES TO BUILD INTERACTIONS



1. Build association rules for all inputs and the target variable
2. Filter out only Consequents that are the target variable.
3. Sort by Confidence or Lift to find strongest rules



USE ASSOCIATION RULES TO BUILD INTERACTIONS



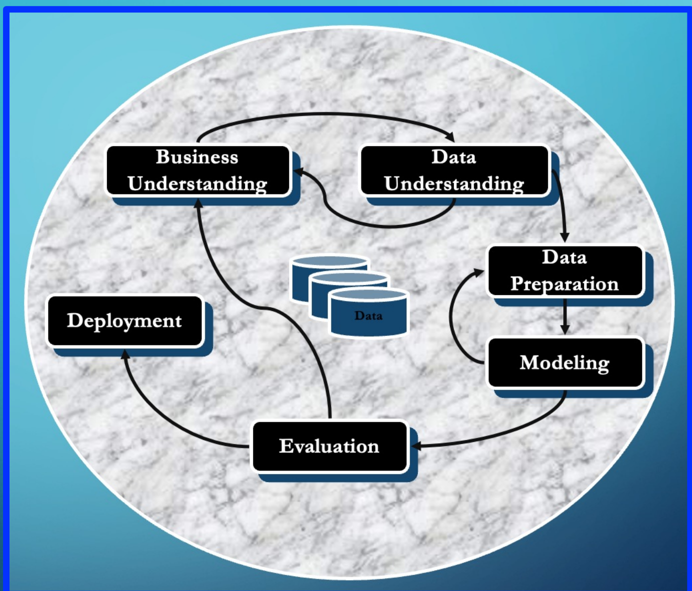
1. Build association rules for all inputs and the target variable
2. Filter out only Consequents that are the target variable.
3. Sort by Confidence or Lift to find strongest rules
4. Note that there may be “duplicate” rules.

Algorithm	Records	Inputs*	Combos	Time
Logistic	95,412	60	1,770	5 hours
AR	95,412	264	34,453	2 minutes

A abbott analytics

A TWELFTH IDEA

12. Don't rely on Generative AI to Do Your Job in Machine Learning.



ChatGPT

The image was generated using DALL-E, an AI model that creates images from text

THE PROMPT FOR CHATGPT 4

- A dataset to analyze with machine learning exists here: [I uploaded the file]. I call it titanic3. It contains 1309 rows and 7 columns. The variable names are in the first row. The target variable is called "survived" and it contains only values 1 and 0, 1 for survivors, and 0 for non-survivors. **Make survived a categorical variable**. The remaining variables can be used as model inputs to predict "survived" equal to 1. The goal is to try to build a machine learning model to predict "survived" most accurately and robustly as possible, **where accuracy means the model that has the highest lift for the top 30% of probability scores**. You may use any algorithm you wish but try at least two different algorithms. Choose the algorithm that has the highest accuracy on hold-out data, but generate a list of the algorithms used to build models.

- If there are **missing values**, replace them with the **mean** and generate a list of every variable and how many missing values the data has had filled in.

- **Please describe the model accuracy on training data and out-of-sample data** you generate from the original data set, including the baseline rate of survived and the lift at the 3rd decile**. Use best practices to determine how to create training and hold-out data sets**. Also**, generate a list of the top 3 strongest predictor variables** along with scatterplots plotting their vs the model predictions on held-out data.

WHAT CHATGPT DID: SAMPLING

What ChatGPT Wrote:

- We'll use a common split ratio of 70% training data and 30% hold-out data, which allows for both robust training and a meaningful evaluation. Let's proceed with these steps.
- The dataset has been split into training and hold-out sets:
- • Training data: 916 samples
- • Hold-out data: 393 samples

WHAT CHATGPT DID: MODEL ACCURACY

What ChatGPT Wrote:

- Logistic Regression Model
  - AUC Score: 0.8646
  - Accuracy at 3rd Decile: 0.7786
- Random Forest Classifier Model
  - AUC Score: 0.8432
  - Accuracy at 3rd Decile: 0.7634

ADDITIONAL PROMPT

- try also to create features to improve model accuracy at the 30% depth**. If any features improve model accuracy on hold-out data**, describe how the feature was created and redo the same reports -- accuracy at the 30% depth and AUC on train and hold-out data, and top 3 variables with scatterplots vs. predicted probability

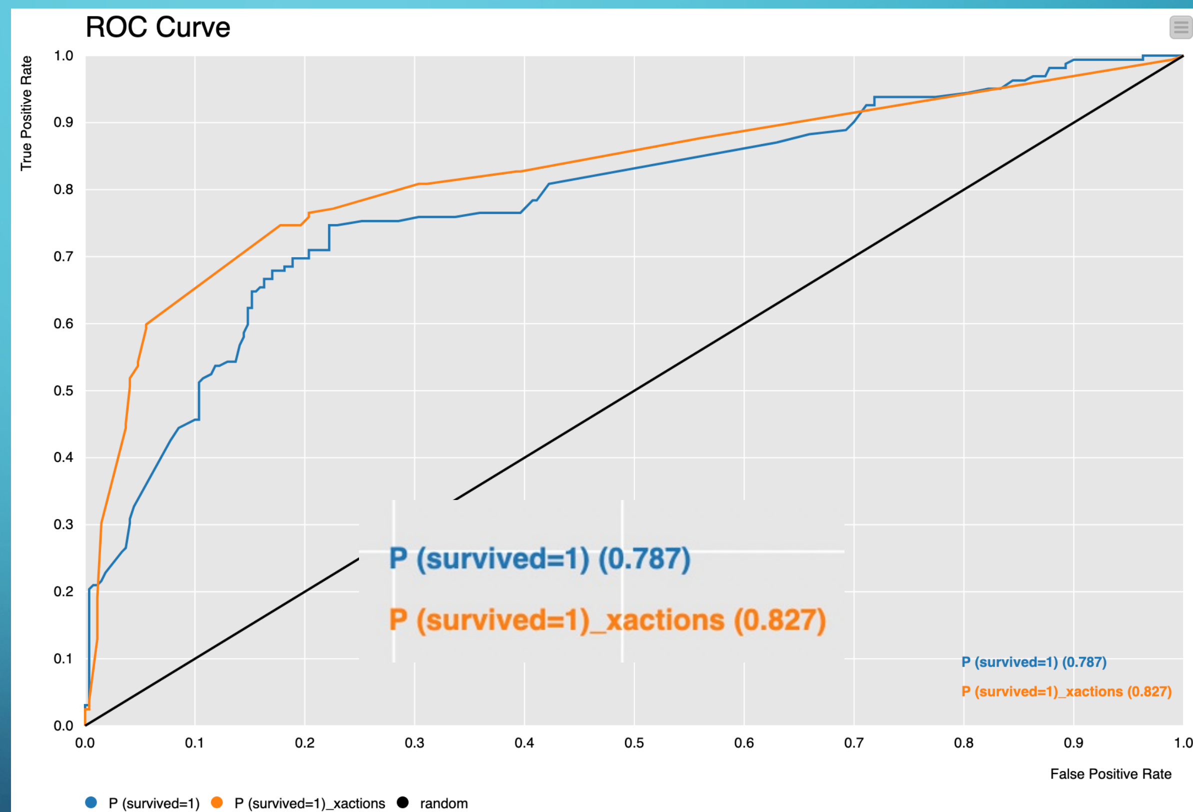
WHAT CHATGPT DID: MODEL ACCURACY

- Let's explore creating some new features from the existing data. Common feature engineering techniques for this type of data might include:
- **Creating interaction features** between pclass and age or fare, to capture the combined effect of socio-economic status with age or fare paid.
- **Binning the age variable** to create categorical age groups, which may help in capturing non-linear age effects.
- **Combining sibsp and parch** to create a 'family size' feature.
- **Creating a feature that signifies whether a passenger was traveling alone or not**.

- The accuracy at the 3rd decile for both extended models has not changed from the original models. This indicates that the additional features did not improve the performance of the models at the 30% depth in predicting survival on the hold-out set.

REMINDER: IMPROVEMENT IN ACCURACY FOR LOGISTIC REGRESSION WHEN INCLUDING 1 INTERACTION TERM

Columns: 20	Column Type
survived	String
age	Number (double)
pclass_1	Number (integer)
pclass_2	Number (integer)
male	Number (integer)
embarked_C	Number (integer)
embarked_S	Number (integer)
cabin_B	Number (integer)
cabin_E	Number (integer)
cabin_D	Number (integer)
cabin_A	Number (integer)
cabin_U	Number (integer)
cabin_F	Number (integer)
sibsp_0	Number (integer)
sibsp_2_3	Number (integer)
sibsp_4+	Number (integer)
parch_0	Number (integer)
parch_2_3	Number (integer)
parch_4+	Number (integer)
fare_log10	Number (double)



Add Feature:
Pclass_2 and female

CHATGPT DID GREAT!
BUT CHATGPT MISSED…BEST INTERACTION TERMS

- Sex interactions
  - Female X pclass
  - Male X age
  - X fare bins

Consequent	Split Value 1	Split Value 2	support_records	Confidence	Lift
survived_true	rules_sex_female	rules_pclass_1	138	0.965	2.527
survived_true	rules_sex_female	rules_(67,513]_fare_true	89	0.918	2.404
survived_true	rules_sex_female	rules_embarked_C	101	0.903	2.363
survived_true	rules_sex_female	rules_pclass_2	93	0.887	2.322
survived_true	rules_(39,80]_age_true	rules_sex_female	69	0.833	2.182
survived_true	rules_sex_female	rules_(26,67]_fare_true	77	0.804	2.105
survived_true	rules_sex_female	rules_(28,39]_age_true	75	0.776	2.030
survived_true	rules_sex_female	rules_[0,8]_fare_false	290	0.756	1.980
survived_true	rules_sex_female	rules_(27,28]_age_false	284	0.755	1.976
survived_true	rules_(13,26]_fare_false	rules_sex_female	257	0.753	1.971

Consequent	Split Value 1	Split Value 2	support_records	Confidence	Lift
survived_false	rules_(8,13]_fare_true	rules_sex_male	166	0.870	1.407
survived_false	rules_[0,8]_fare_true	rules_sex_male	239	0.861	1.393
survived_false	rules_sex_male	rules_pclass_2	145	0.854	1.381
survived_false	rules_sex_male	rules_pclass_3	414	0.848	1.372
survived_false	rules_(26,67]_fare_false	rules_sex_male	568	0.840	1.359
survived_false	rules_(39,80]_age_true	rules_sex_male	134	0.839	1.357

CONCLUSIONS

- Incredible Growth in Data Science and Machine Learning
  - Depth *and* Breadth of Education in This Generation of Data Scientists
- There's More to Being a Data Scientist than Just the "Science"
- Keep the Goal of the Analysis in Mind
- Understand the Limits and Weaknesses of the Science
- Think Before You Deploy!

感謝你
Danke
Gracias
Grazie
Merci
Thanks!
THANK YOU!

WHAT DATA SCIENTISTS DISCOVER THEY ACTUALLY DO

Business Understanding

1. Define the Target Variable Carefully and Correctly
2. Connect the measure of model success to the model's Business Objective.

Data Preparation

3. Account for Dummy Variables appropriately (differently from Other Numerical Variables)
4. Create Interaction Features at Scale to Improve Machine Learning Algorithm Accuracy
5. Beware of the Destructive Influence of Irrelevant Variables In Models

Modeling

6. Remember that Even Robust Ensembles Can Overfit
7. Account for High Class Imbalance with the Algorithm in Mind
8. Beware of Ways Even Advanced Algorithms are Fooled By Data

Evaluation

9. **Be Prepared to Quantify Machine Learning Model Prediction Stability.**
10. Quantify Variable Importance Related to Its Use When Deployed

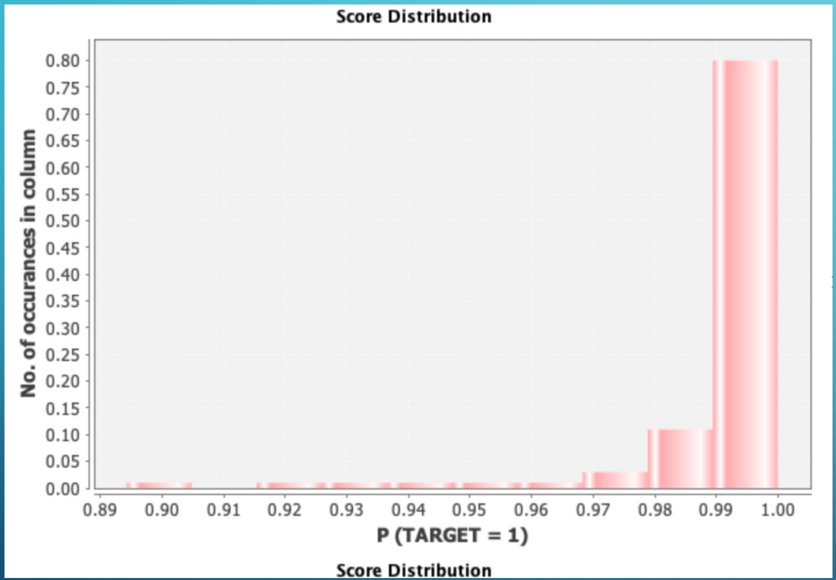
Deployment

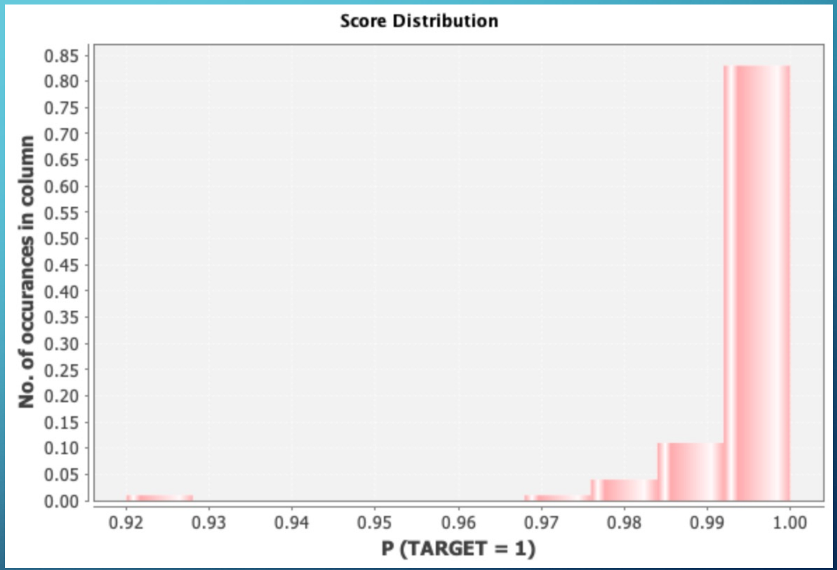
11. Prepare for Model Obsolescence: When and How to Rebuild Models

THE APPROACH

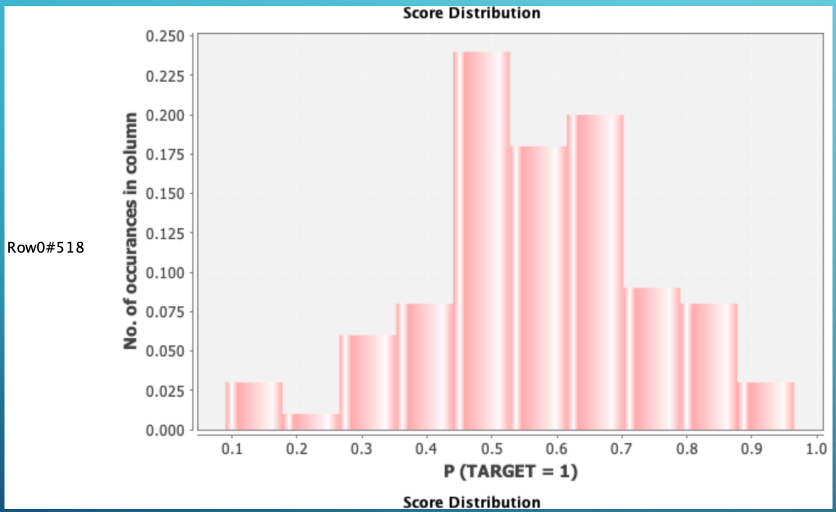
- Bootstrap sample the training data (~100x)
- Build a model on each sample
- Compute the prediction for each record
- Compute the range / standard deviation for each record (100 models)
  - Small range or standard deviation means every model predicts a similar value -> consistent predictions
  - Large range or standard deviation means models disagree, and the record isn't

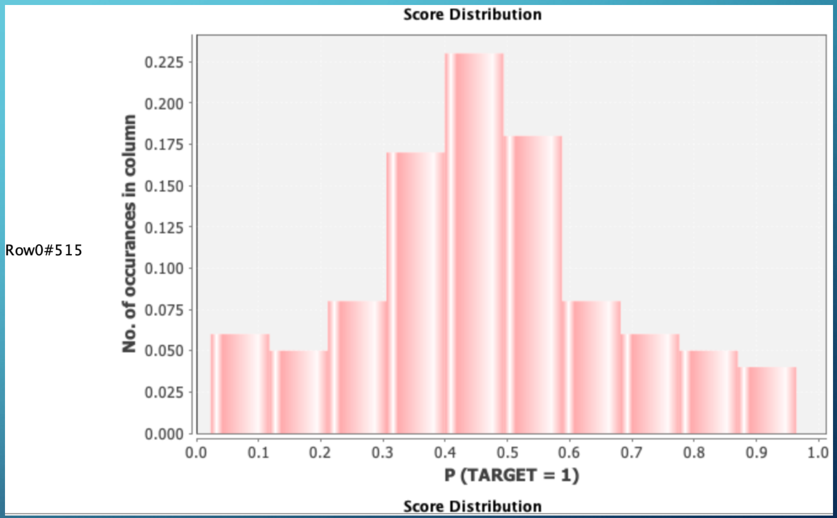
FOR SOME RECORDS, CONFIDENCE IN THE PREDICTED SCORE IS STABLE





CONFIDENCE IN THE PREDICTED SCORE CAN VARY WILDLY





A FEW EXAMPLES: LOW CONFIDENCE

Measure	record 1	record 2	record 3	record 4
Count(TARGET)	100	100	100	100
Mean(P (TARGET=1))	0.959	0.466	0.518	0.723
Standard deviation(P (TARGET=1))	0.112	0.329	0.318	0.272
Min*(P (TARGET=1))	0.003	0.000	0.000	0.000
Max*(P (TARGET=1))	1.0000	0.9999	0.9500	1.0000
Range(P (TARGET=1))	0.9966	0.9998	0.9500	1.0000
survived	1	0	1	1
age	53	24	37	25
pclass_1	1	1	1	1
pclass_3	0	0	0	0
female	1	0	0	0

A FEW EXAMPLES: HIGH CONFIDENCE

Measure	record 1	record 2	record 3	record 4
Count(TARGET)	100	100	100	100
Mean(P (TARGET=1))	0.995	0.994	0.988	0.993
Standard deviation(P (TARGET=1))	0.010	0.014	0.018	0.015
Min*(P (TARGET=1))	0.942	0.912	0.921	0.901
Max*(P (TARGET=1))	1.0000	1.0000	1.0000	1.0000
Range(P (TARGET=1))	0.0580	0.0881	0.0791	0.0992
survived	1	1	1	1
age	32	22	64	17
pclass_1	1	1	1	1
pclass_3	0	0	0	0
female	1	1	1	1