A Brief Tour of Data Science Statistical Methods

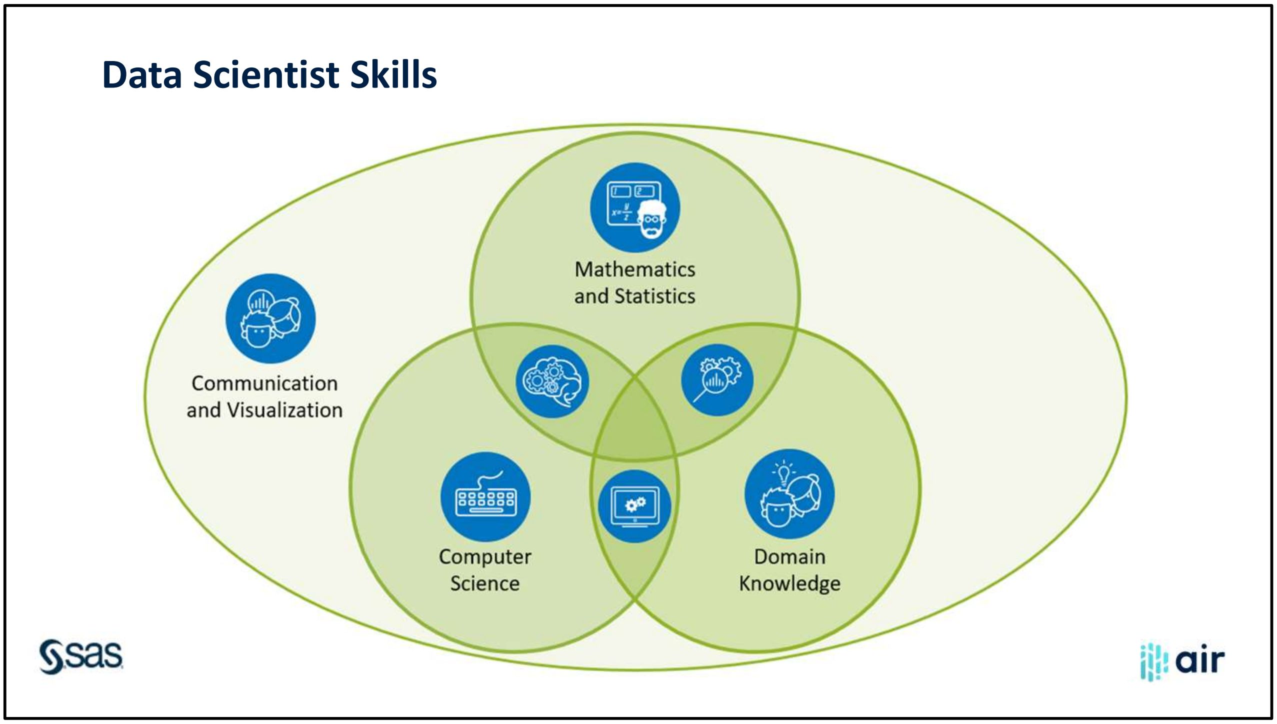
Jacqueline Johnson
Principal Analytical Training Consultant
SAS Institute

A Brief Tour of Data Science Statistical Methods
The field of data science includes contributions from a diversity of disciplines including computer science, statistics, and machine learning. In this webinar we will describe the different skill domains important for data scientists, with focus on machine learning models. We will walk through the steps of data exploration, data preparation, and statistical modeling conducted in the data science model building process and examine some of the most commonly used supervised and unsupervised machine learning models.

As a result of this webinar, participants will be able to:

- Describe the goals of machine learning models
- Identify some commonly used machine learning models

Presenter: Jacqueline Johnson, Principal Analytical Training Consultant, SAS Institute



Data Scientist Skills

Data science skills are quite varied. Some of the necessary skills are mathematics and statistics. Data scientists need mathematics and statistics to understand modeling, modeling techniques, and model evaluation.

Because data scientists do not develop models by hand, they also need to have computer science skills to develop code as well as to extract, prepare, merge, and store data. With the increasing popularity of cloud-based computing, data scientists need to know how to put data models into the cloud and use containers. A container packages up software code and all its dependencies, so the application runs quickly and reliably from one computing environment to another.

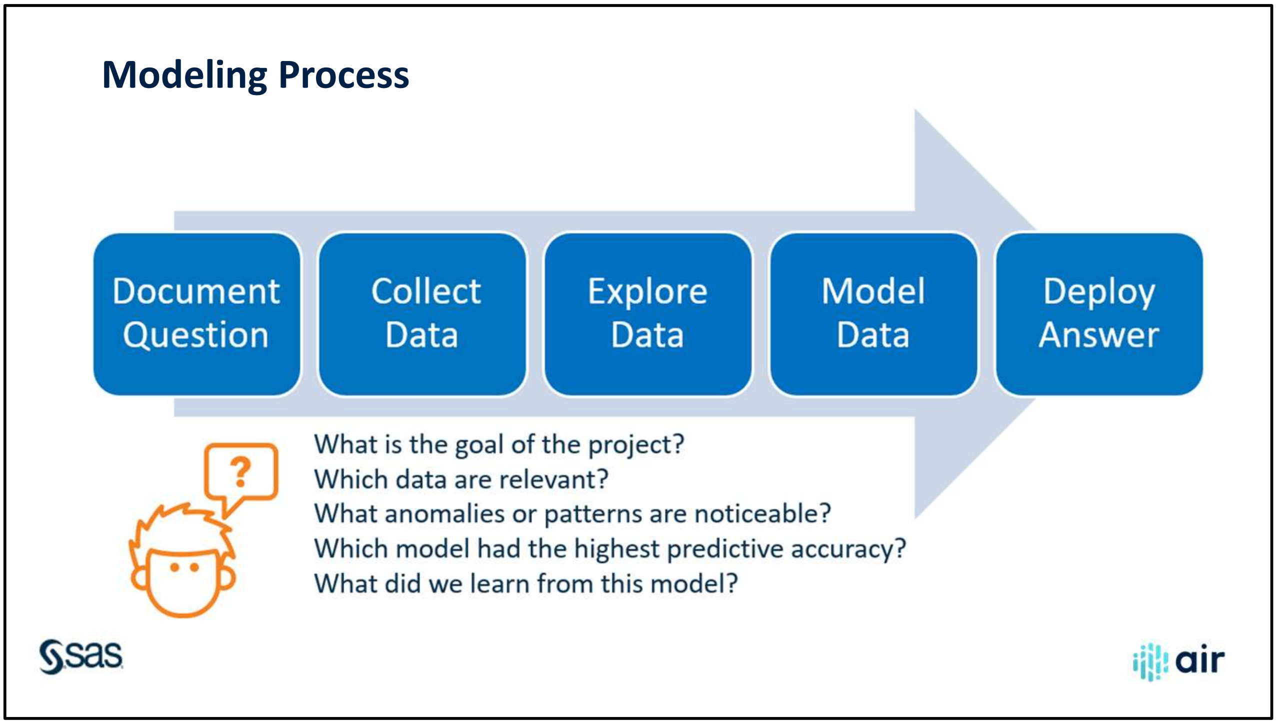
One of the most critical skills data scientists need is domain knowledge. Knowing the business problem helps data scientists create new predictor variables (also called feature engineering). In the telecommunications field, for example, data scientists need to know what type of data are available, how transaction systems are implemented, what billing system is used, and how data from the call center are collected.

Machine learning, where mathematics, statistics, and computer science meet, is a method of data analysis that automates analytical model building. Machine learning enables data scientists to implement complex models, such as neural networks.

Research skills that enable data scientists to apply new techniques in model building come into play where mathematics and statistics combine with domain knowledge.

Software skills, like SAS, help data scientists create new models and fully use both computer science skills and domain knowledge.

Finally, data scientists need communication and visualization skills to be able to explain models, interpret models, and graphically illustrate results of models. Results of models can be used to create marketing campaigns, offer insights into customer behavior, and lead to business decisions and actions.



Modeling Process

Document Question
Collect Data
Explore Data
Model Data
Deploy Answer

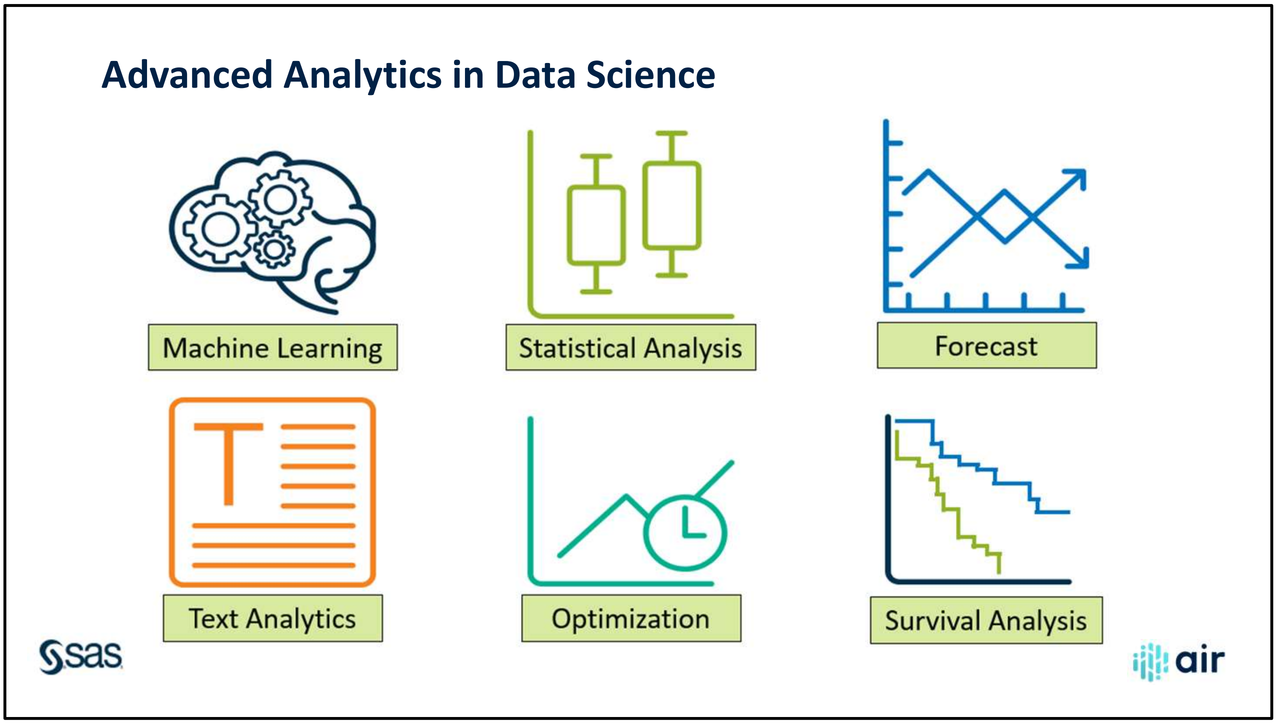
What is the goal of the project?
Which data are relevant?
What anomalies or patterns are noticeable?
Which model had the highest predictive accuracy?
What did we learn from this model?

The first step in the modeling process is to document the question. Data scientists need to understand what they are trying to solve with this model.

The next step in the data science modeling process is to collect the data. Some desired variables, such as gender and income, for example, might not be available to use even though those predictors might be associated with the outcome. Furthermore, even if the data are available when you develop the model, will the data be available when the model is in production, such as scoring a business transaction for fraud?

The next step is to explore the data. In some projects, for example, data scientists might have thousands of predictor variables, which is far too many to model. Data scientists might also have several predictor variables with missing values, which will need to be replaced with reasonable values.

The next step is modeling the data. This is the data scientists' playground, where they use different algorithms and techniques. In this phase, they fit the model on a portion of the data and evaluate the model's performance on another part of the data. It should be noted that sometimes the best model is the most interpretable and simplest model rather than the one with the highest predictive accuracy.



Advanced Analytics in Data Science

Machine Learning

Statistical Analysis

Forecast

Text Analytics

Optimization

Survival Analysis

Data science encompasses more than simply statistical analysis. The field encompasses machine learning, forecasting, text analytics, optimization, and survival analysis. Data scientists use all these techniques to solve business problems.

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

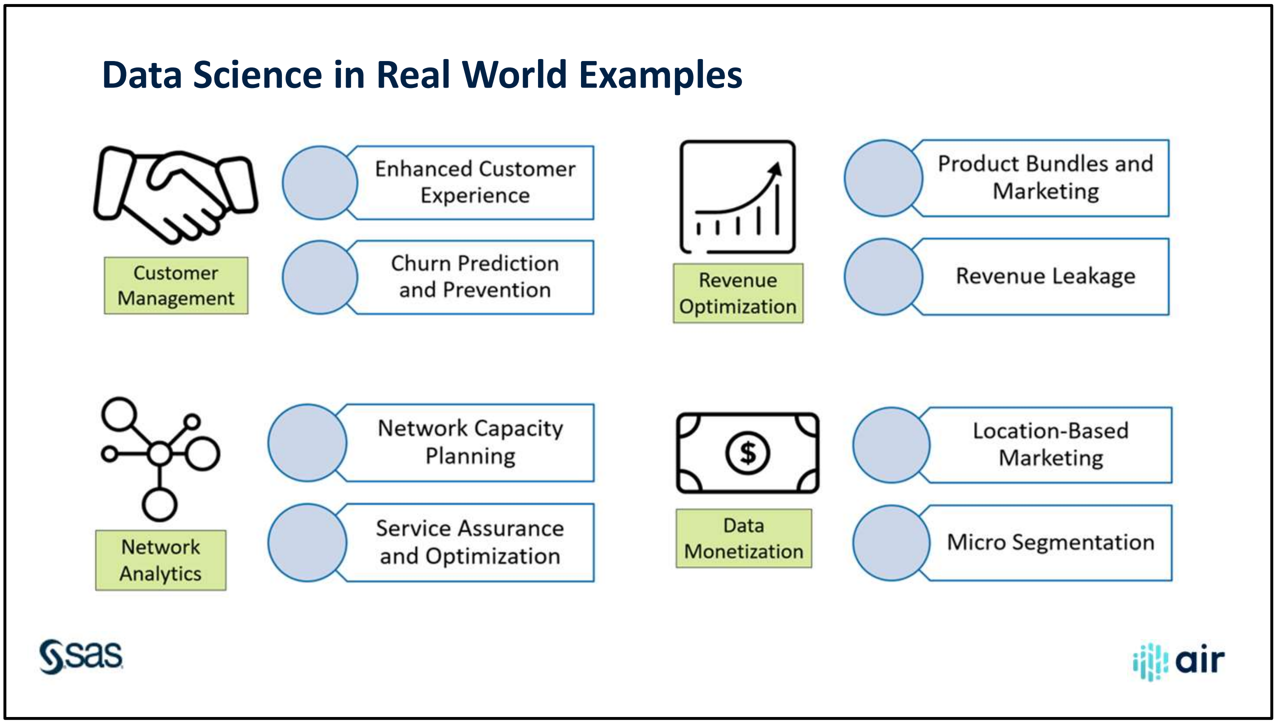
Statistical analysis is the science of collecting, exploring, and presenting data to discover underlying patterns and trends in spite of the inherent variability in the data. Companies use statistics every day to make informed business decisions. The statistical analysis field includes descriptive statistics, where data scientists summarize data to describe the past. Another statistical analysis field is inferential statistics, where data scientists take the descriptive statistics from a sample and generalize them to a larger population.

Forecasting describes an observed time series to understand the underlying causes of changes and to predict future values. It involves assumptions about the form of the data and decomposing the time series into components.

Data scientists use text analytics to uncover insights hidden in text data with the combined power of natural language processing, machine learning, statistics, and linguistic rules. They use text analytics to analyze unstructured text, extract relevant information, and transform it into useful business intelligence (for example, data scientists use information retrieval to find material of an unstructured nature, such as using a search engine to find specific material).

Optimization is a major component of operations research, industrial engineering, and management science. Simply put, optimization is the process of choosing the available actions that produce the best results. It goes beyond “what happened”, which is answered in descriptive statistics, and “what will happen”, which is answered in predictive statistics. Optimization addresses the question “what should you do”.

Survival analysis is a class of statistical methods for which the outcome variable of interest is time until an event occurs. Time is measured from when an individual first becomes a customer until the event occurs or until the end of the observation interval (the individual then becomes censored). In survival analysis, the basis of the analysis is tenure, or time at risk for the event. Therefore, it is not just whether the event occurred, but when it occurred. The goal in survival analysis is to model the distribution of time until an event.



Data Science in Real World Examples

Customer Management

- Enhanced Customer Experience
- Churn Prediction and Prevention

Revenue Optimization

- Product Bundles and Marketing
- Revenue Leakage

Network Analytics

- Network Capacity Planning
- Service Assurance and Optimization

Data Monetization

- Location-Based Marketing
- Micro Segmentation

Data science can help companies address business challenges. Some examples include:

Enhanced Customer Experience – Data scientists can enhance customer experience by offering more personalized and relevant customer marketing promotions.

Churn Prediction and Prevention – Data scientists can forecast customer issues and prevent them from happening (for example, if a utility company forecasts an outage and sends a text to the customers affected by the outage, this might improve the customer experience).

Product Bundles and Marketing Campaigns - Data scientists can improve business decision making processes related to product bundles and marketing campaigns.

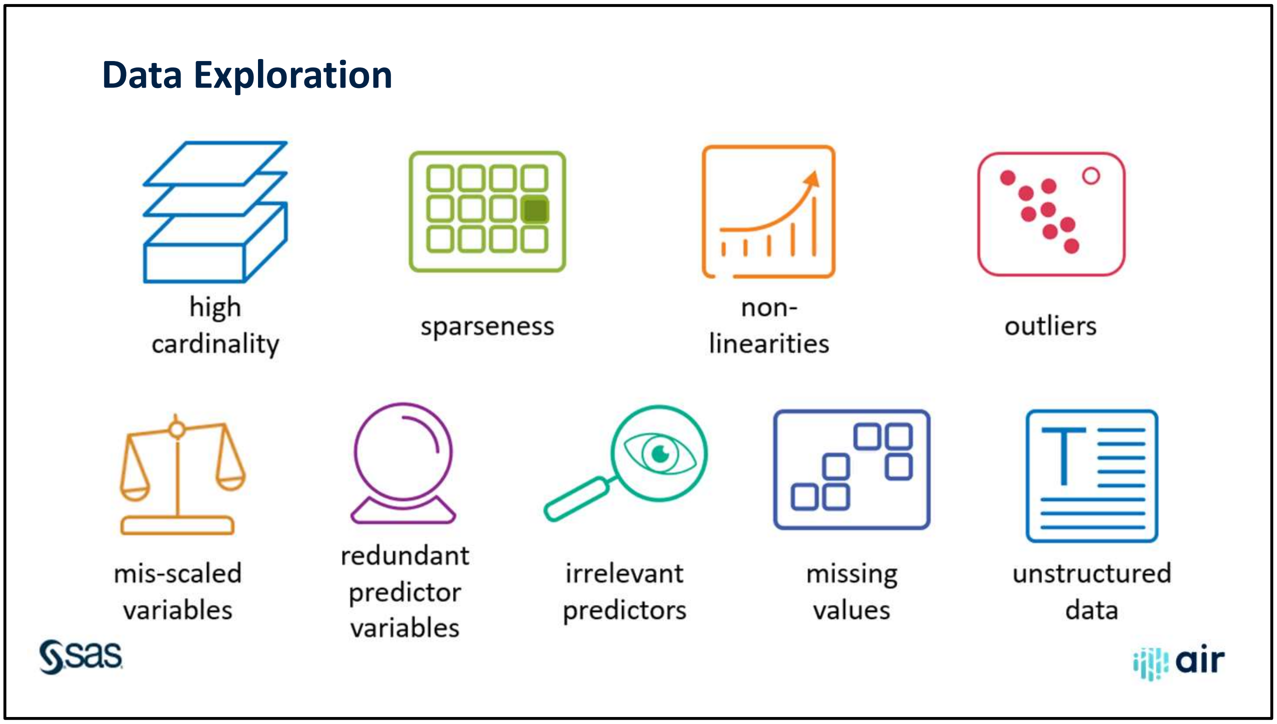
Revenue Leakage – Data scientists can identify situations leading to revenue leakage, whether due to billing and collections, network, or fraud issues (for example, a collection agency might be interested in who is more likely to pay their debts rather than who owes the most money).

Network Capacity Planning – Data scientists can use statistical forecasting and detailed network or supply chain data to more accurately plan capacity (for example, they can use network analytics on ATM machines to make sure they have enough cash to meet customer demands).

Service Assurance and Optimization – Data scientists can use network analytics to prevent network or supply chain problems before they happen.

Location-Based Marketing – Data scientists can develop specialized offers and promotions that are delivered to targeted customers via their mobile devices (for example, if you enter a specific area in a city, you might get a coupon for Starbucks or McDonalds).

Micro Segmentation – Data scientists can create highly detailed customer segments that they can use to send highly targeted and timely mobile messages.



Data Exploration

high cardinality

sparseness

non-linearities

outliers

mis-scaled variables

redundant predictor variables

irrelevant predictors

missing values

unstructured data

One of the first steps in the modeling process is data exploration, when data scientists get to know the data.

Some of the variables in the data might have high cardinality, meaning that categorical variables have numerous levels (for example, variables such as ZIP codes or product codes might have high cardinality).

The data scientists might also encounter sparseness in the data, where there are very few events in the data.

Data exploration can also illuminate non-linear associations between the predictor variable and the outcome. In this situation, the data scientists modify the model to take the non-linearities into account.

Outliers can also be detected during this phase. The data scientists decide whether these data points are erroneous or depict unusual circumstances.

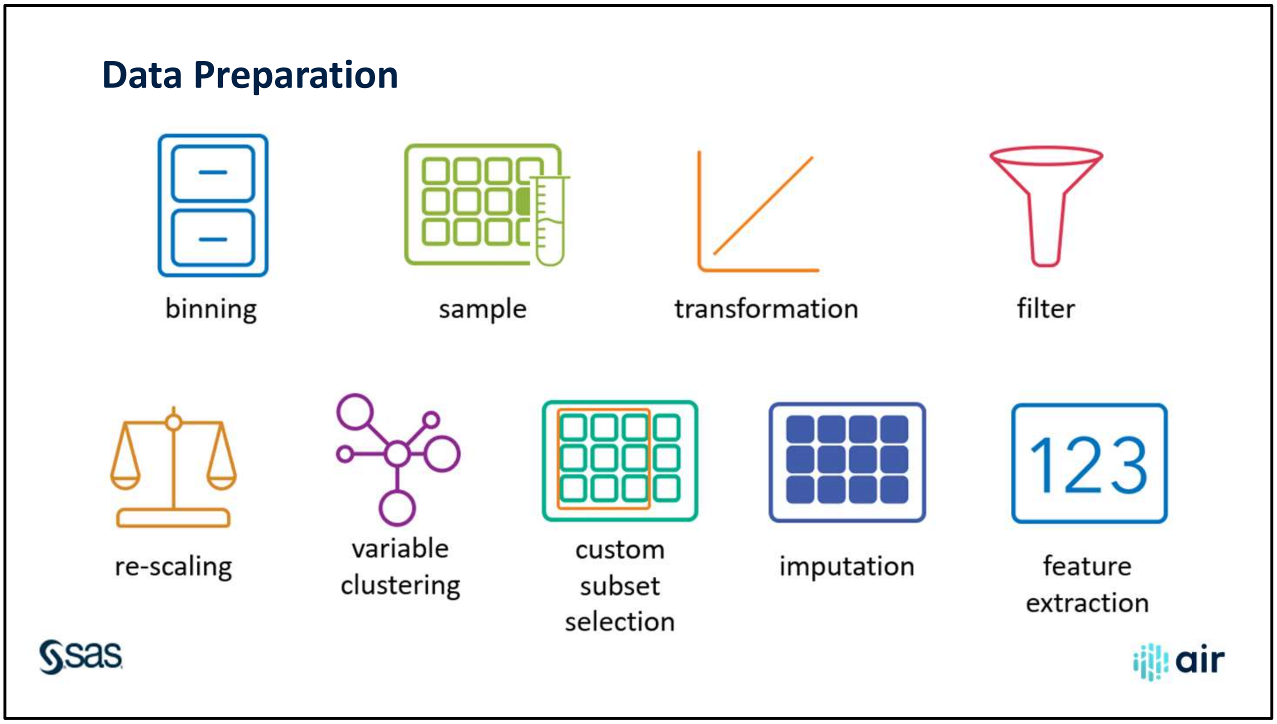
Data scientists should examine the scale of the predictor variables and might need to address mis-scaled variables.

Identifying redundant predictor variables, where the variables have similar information, is an important step during data exploration. Data scientists try to reduce the number of redundant variables to a subset of relatively independent variables for model stability.

Examining the relationship between the predictor variables and the outcome is useful for identifying irrelevant predictors. If the data scientists have several hundred predictor variables, reducing the number of redundant and irrelevant variables to a reasonable subset of variables in the data exploration phase of model building is a useful practice.

Identifying variables with missing values allows data scientists to impute missing values with a reasonable value.

Examining unstructured data, such as textual data, network data, and image recognition data, can be helpful, as unstructured data might be good predictors of the outcome (for example, in a churn analysis, examining call center data and identifying unhappy customers can lead to good predictors for churn).



Data Preparation

binning

sample

transformation

filter

re-scaling

variable clustering

custom subset selection

imputation

feature extraction

In the data preparation phase, data scientists apply remedies to problems identified in the data exploration phase.

If there are variables with high cardinality, data scientists bin the variable into a few groups (for example, data scientists might bin ZIP codes into a few geographic regions).

If there is sparseness in the data, data scientists might want to sample the data.

Data scientists can transform the predictor variable if there are non-linear associations between the predictor variable and the outcome.

Filtering the outliers and replacing the erroneous values with accurate information enables data scientists to address outliers in the data.

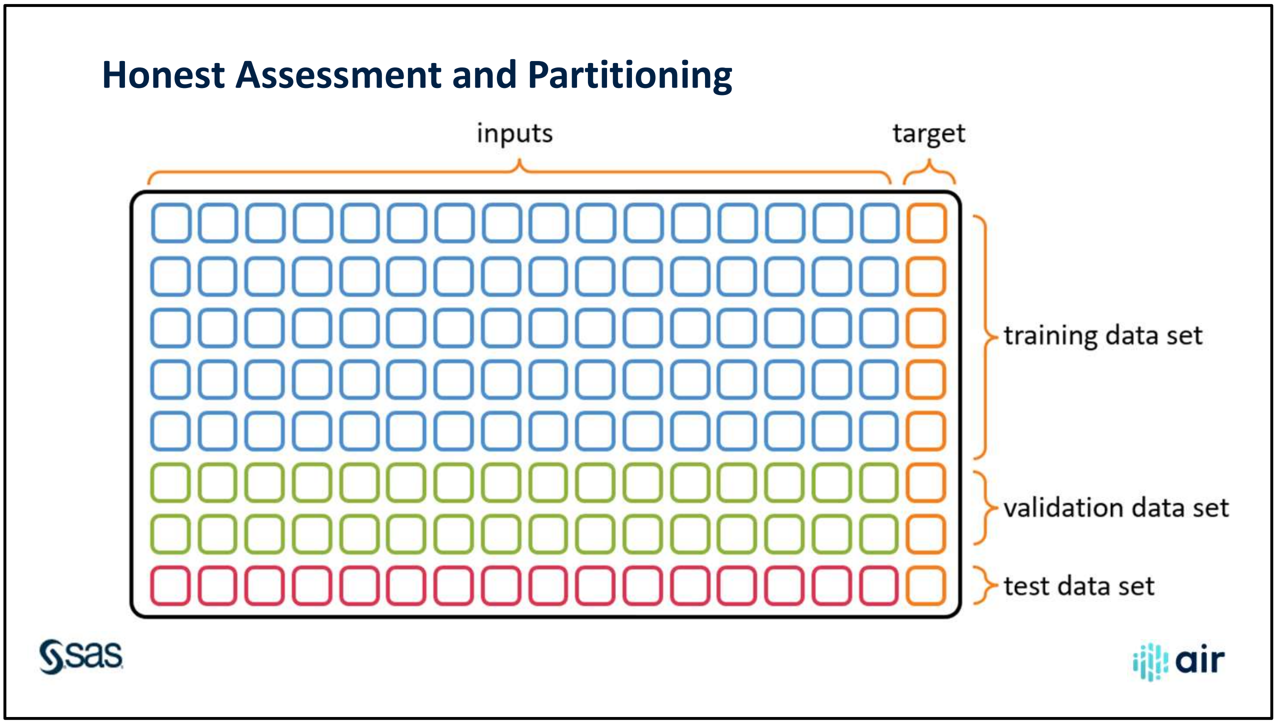
In some modeling situations, re-scaling the variables to the same scale might be beneficial.

Data scientists might perform variable clustering and choose one variable from each cluster if there are redundant predictor variables.

On the other hand, if there are many irrelevant variables, data scientists might perform custom subset selection to identify the potentially relevant variables.

If there are missing values, data scientists can impute the missing data.

If there are unstructured data, data scientists can perform feature extraction to create new variables from the data.



Honest Assessment and Partitioning

inputs

target

training data set

validation data set

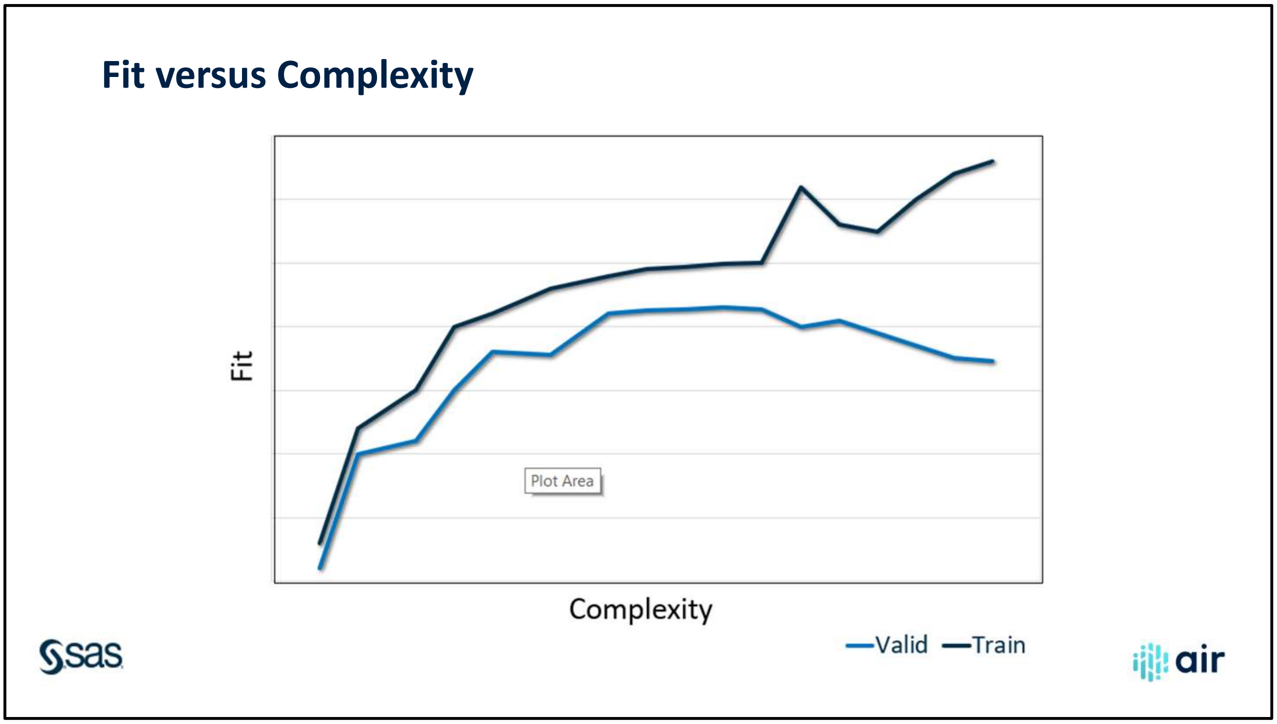
test data set

Evaluating the model on the data the model was fit on usually leads to an optimistically biased assessment. The simplest strategy for correcting the optimism bias is to isolate a portion of the data for assessment.

The model is fit to one part of the data, called the training data set.

The performance is then evaluated on another part of the data, called the *validation data set*.

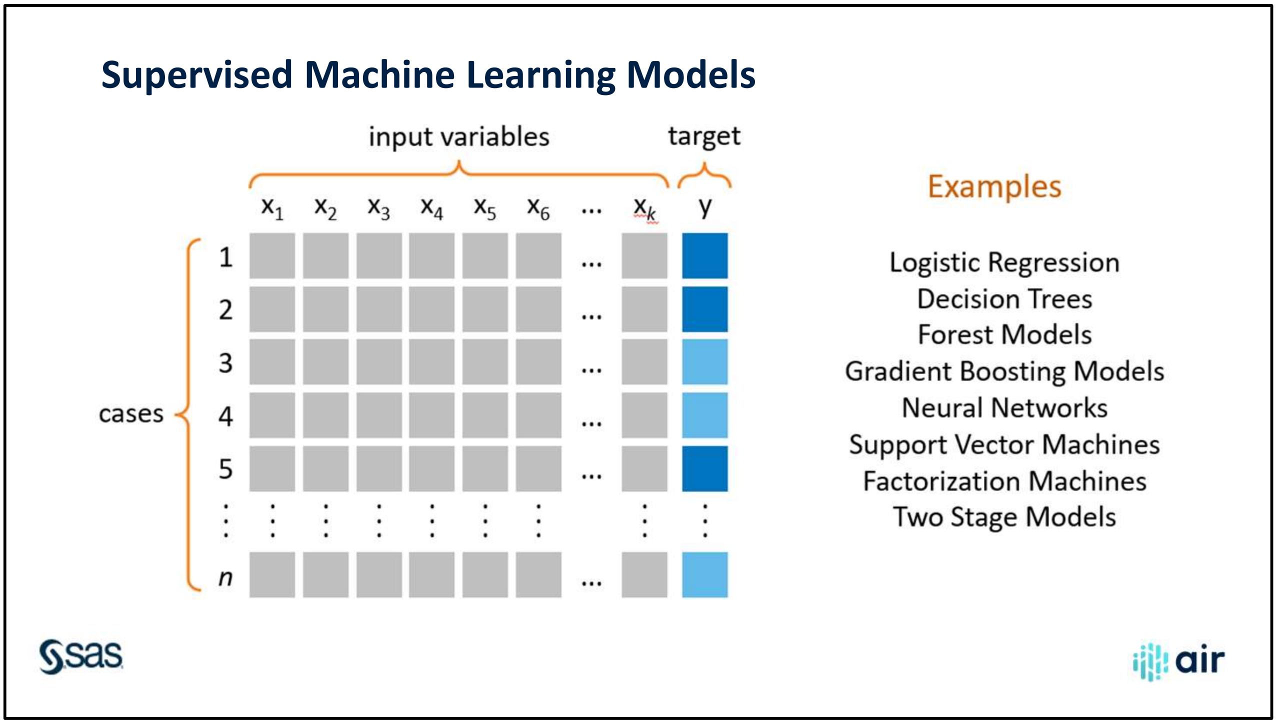
A third test data set can optionally also be used for a final assessment. In situations where there is a time component, the test data set could be gathered from a different time (for example, a model that is fit on data that is gathered from January to June might not generalize well to data that was gathered from July to December).



Fit versus Complexity

Typically, model performance follows a straightforward trend. As the complexity of the model increases and more terms are added to the model, the fit on the training data set generally improves. Some of this increase is attributed to the model capturing relevant trends in the data. However, some of the increase might be due to overfitting as the model reacts to random noise. Therefore, data scientists examine the model fit on the validation data for models of varying complexity.

Typically, the model fit on the validation data increases as the complexity increases, followed by a plateau, followed by a decline in performance. This decline is due to overfitting. Consequently, it is recommended that data scientists select a model that is associated with the complexity that has the highest validation fit statistic.



Supervised Machine Learning Models

Examples

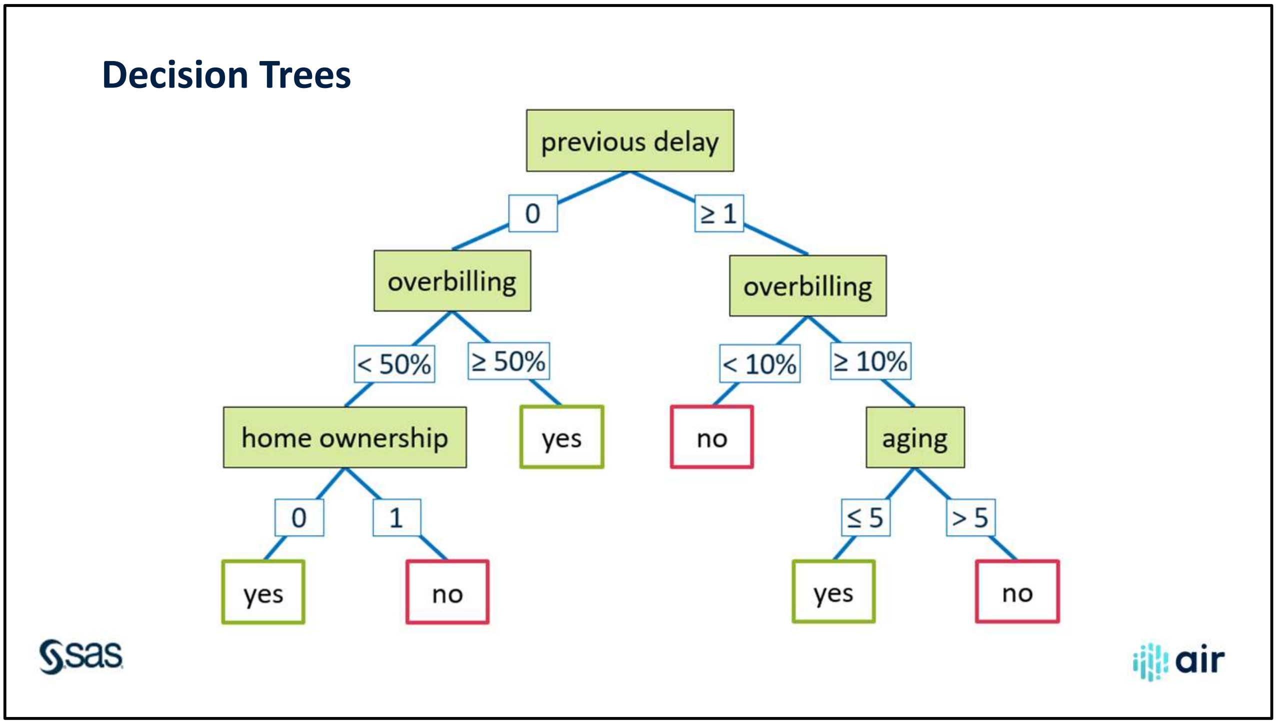
Logistic Regression
Decision Trees
Forest Models
Gradient Boosting Models
Neural Networks
Support Vector Machines
Factorization Machines
Two Stage Models

When the target variable value is known for each case, the machine learning model is called supervised. However, if the target variable value is known, then why build a predictive model?

The principal aim of supervised machine learning models is generalization, or the ability to predict the outcome on new cases.

Generalization is also involved in model assessment. The model is fit to the training data set, and the performance is evaluated on the validation data set by comparing the predicted values to the observed values of the target.

The data used to develop a supervised machine learning model consists of a set of cases, which are also known as observations or examples. Each case is associated with a vector of input variables, which are also referred to as predictors, explanatory variables, and features. Each case also has a target variable, also called an outcome or response. A machine learning model maps the vector of input variables to the target. The target is the outcome to be predicted. The cases are the units on which the prediction is made.



Decision Trees

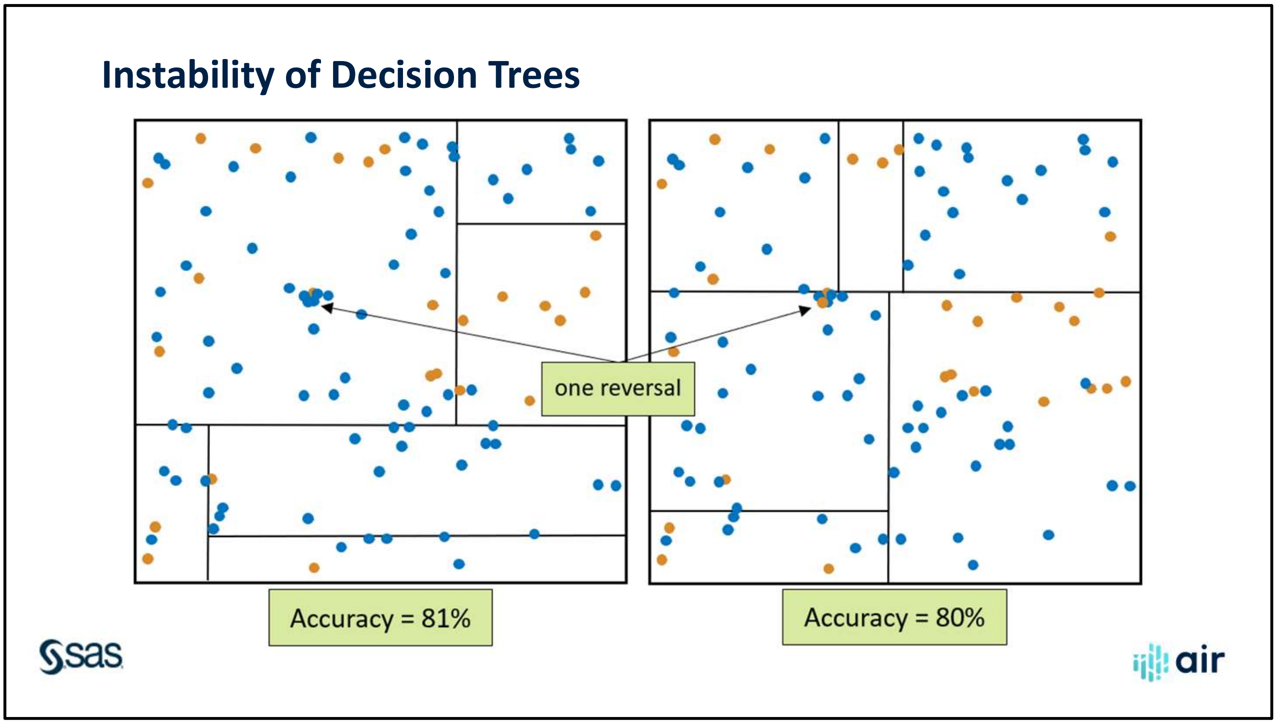
Decision trees are statistical models designed for supervised prediction problems. Cases are scored using a sequence of prediction rules.

To illustrate decision trees using business data, a telecommunications data set is used. The target variable is account default (yes/no). The predictor variables are Previous Delay, Overbilling, Home Ownership (yes/no), and Aging. Previous Delay is the number of previous delays as of January 1, 2020. Overbilling is the billing amount / average billing amount. Aging is the time since the customer first purchased a product in years.

The model is called a decision tree because it can be represented in a tree-like structure. A decision tree is read from the top down starting at the root node. In this example, previous delay is the first variable in the tree and is used in the root node. Cases with a value of 0 go to the left in the tree and cases with values of 1 or more go to the right in the tree. The predictor variable previous delay was chosen for the root node based on a split-search algorithm which finds the predictor variable that gives the most significant split between the values of the target variable.

Each internal node represents a split based on the values of one of the predictor variables. The predictor variables can appear in any number of splits throughout the tree. Cases move down the branch that contains its predictor value.

The terminal nodes of the tree are called leaves. The leaves represent the predicted target. All cases reaching a leaf are given the same predicted value. This value is based on the target value of cases in the training data that reach this leaf. In this example, customers with previous delays in payments, with a billing difference of 10 percent and greater, and who have been customers less than or equal to 5 years are predicted as a default.



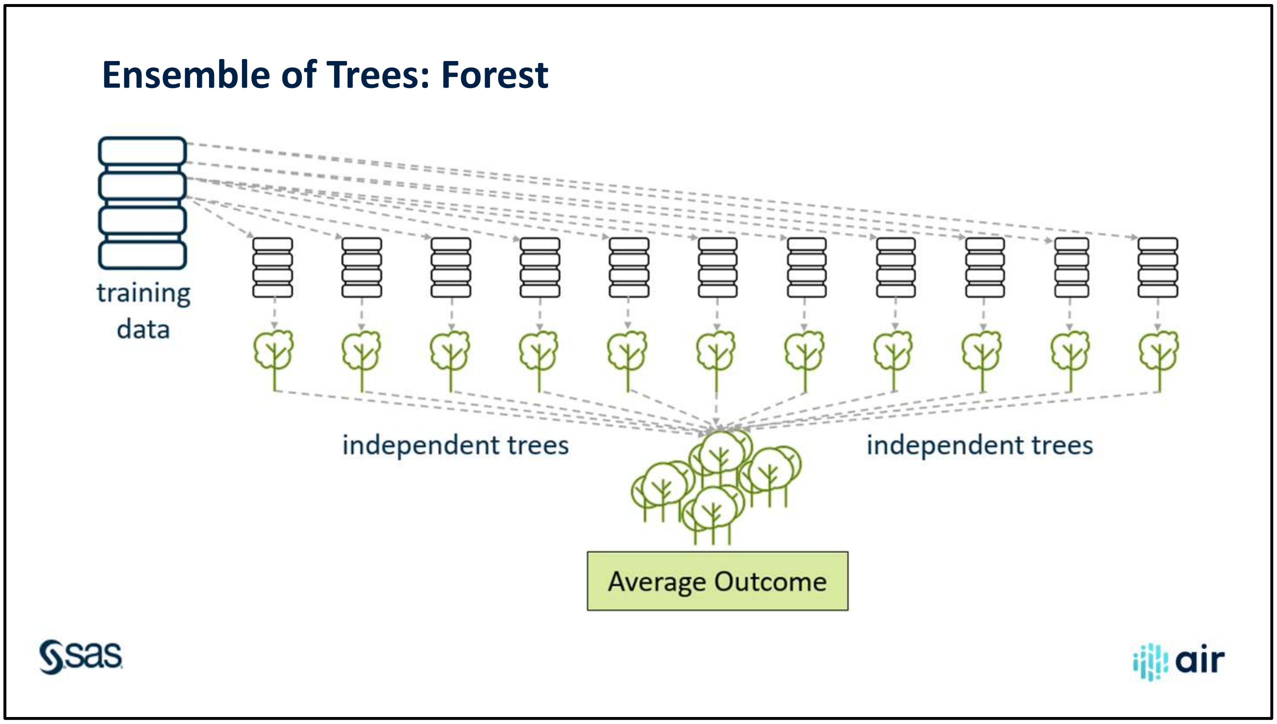
Instability of Decision Trees

one reversal

Accuracy = 81%

Accuracy = 80%

Decision trees are unstable models. Changing the class label of one case could result in a completely different tree with nearly the same accuracy. The instability results from the large number of univariate splits and the fragmentation of the data. At each split, there are typically many splits on the same predictor variable or different predictor variables that give similar performance. For example, suppose income is split at 50,000 since it is the most significant split with the predictor variable and the target. However, other splits at 51,000 or 52,000 might be almost as significant. A small change in the data can easily result in an effect that can cascade and create a totally different tree.



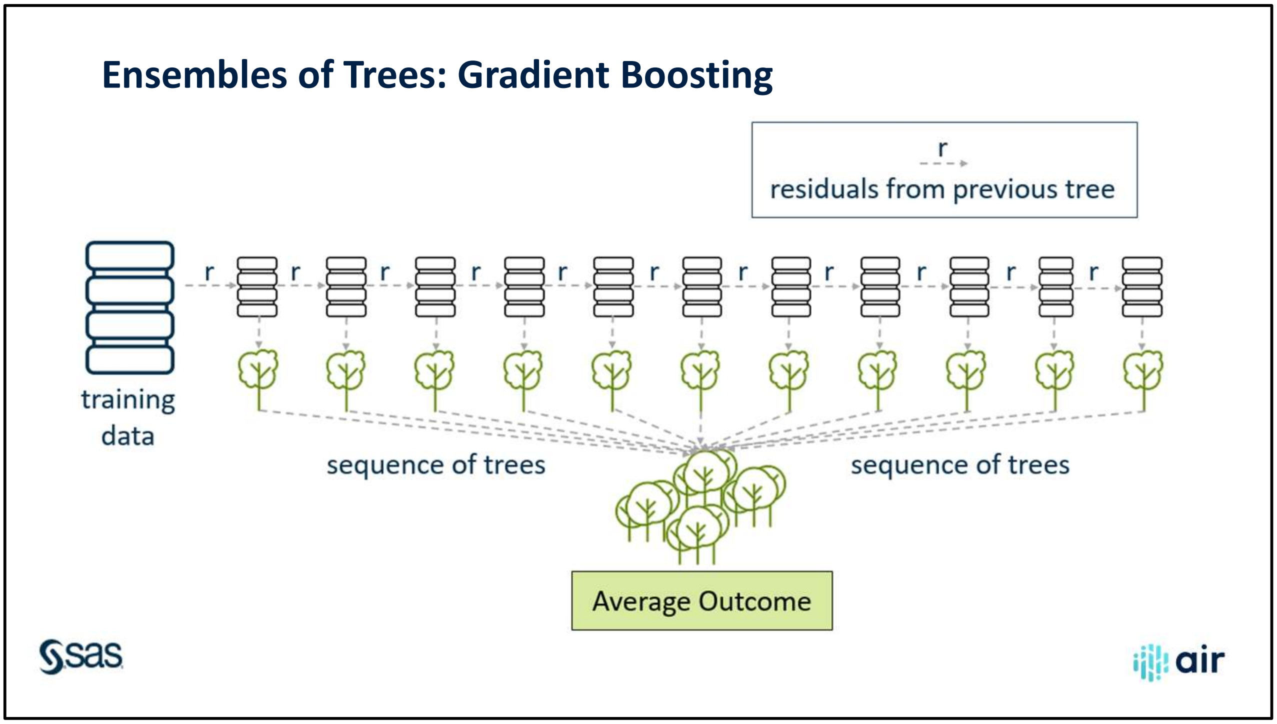
Ensemble of Trees: Forest

Methods have been devised to take advantage of the instability of decision trees to create models that are more powerful. One such model is the forest model which is an ensemble of decision trees.

Decision trees that make up the forest differ from each other in two ways. First, the training data for each tree are sampled with replacement from all observations that were originally in the training data set. Sampling with replacement means that the customer who was sampled is returned to the training data set before the next customer is sampled. Also, the predictor variables considered for splitting for any given decision tree are randomly selected from all available predictor variables.

Therefore, each decision tree is created on a sample of predictor variables and from a sample of the cases. Repeating this process many times leads to greater diversity in the trees. The final model is a combination of the decision trees where the predicted values are averaged.

Forest models usually have improved predictive accuracy over the single decision trees because of variance reduction. If the single decision trees have low bias but high variance, then averaging them decreases the variance.



Ensembles of Trees: Gradient Boosting

r

residuals from previous tree

training data

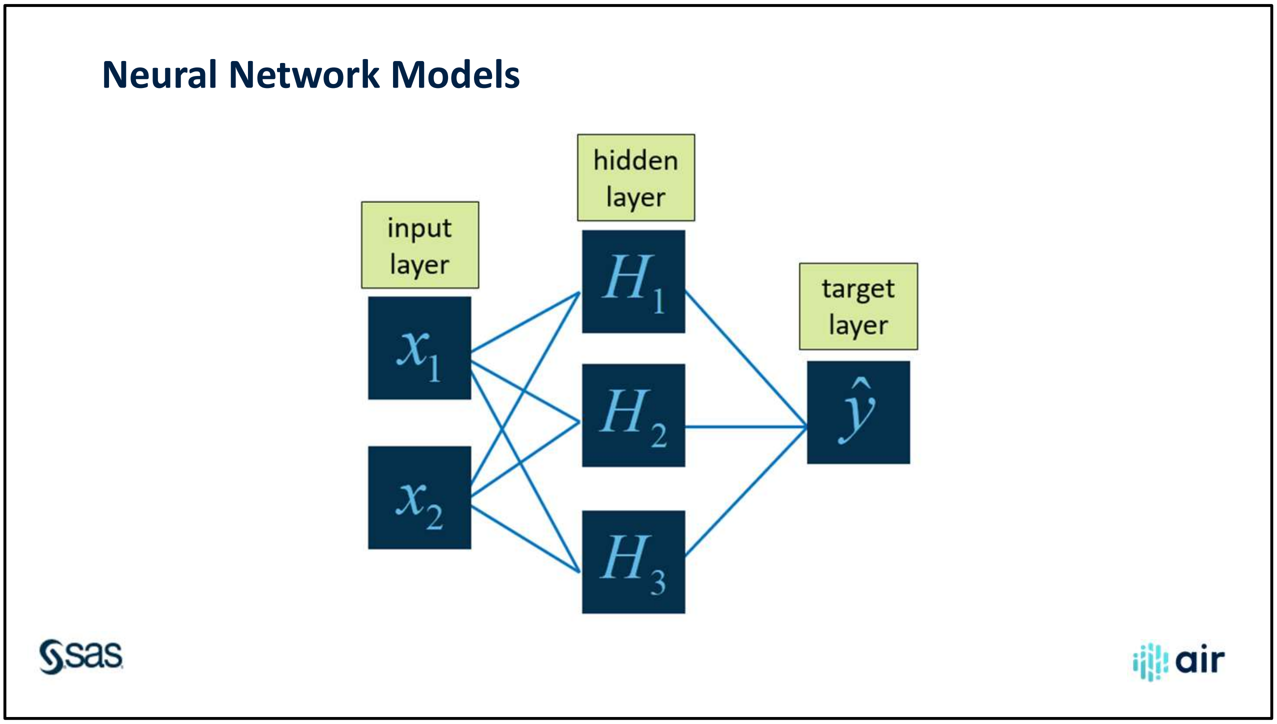
sequence of trees

sequence of trees

Average Outcome

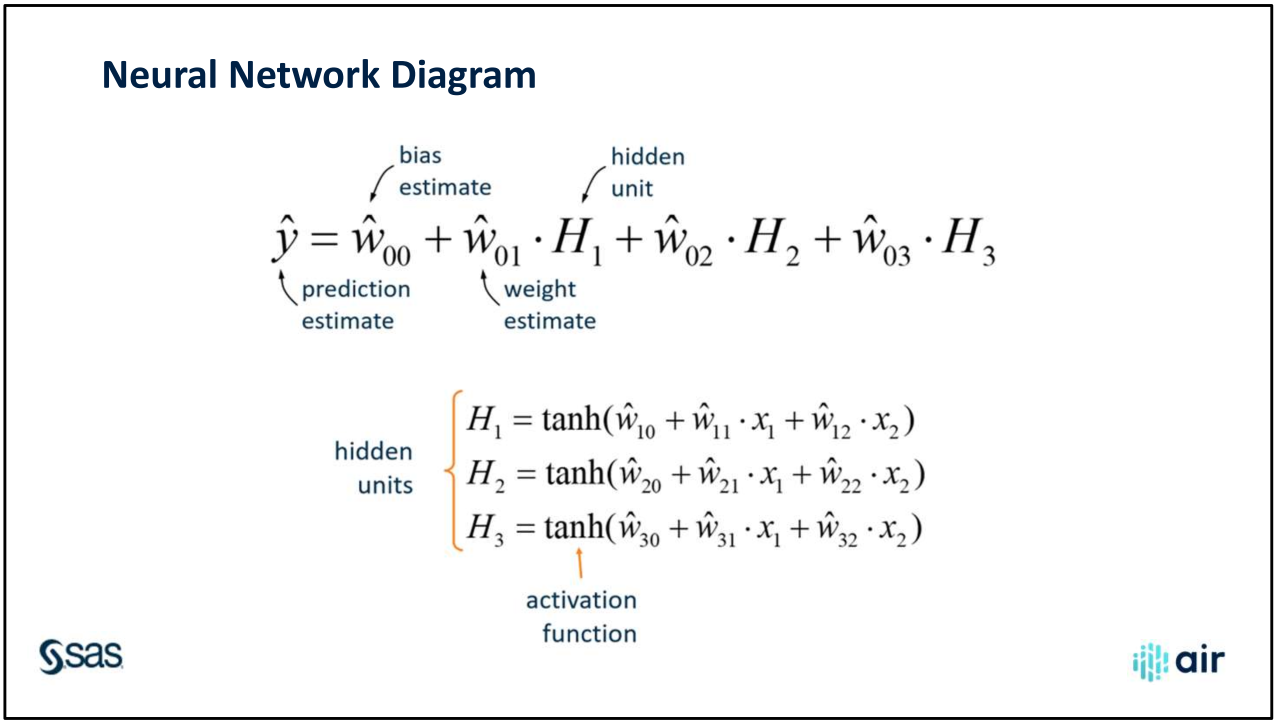
Another model that is a combination of decision trees is the gradient boosting model, which is a weighted linear combination of decision trees. The algorithm starts with an initial decision tree and generates the residuals. In the next step the target is the residuals from the previous decision tree. At each step the accuracy of the tree is computed, and successive trees are adjusted to accommodate previous inaccuracies. Therefore, the gradient boosting algorithm fits a sequence of trees based on the residuals from the previous trees. The final model also has the predicted values averaged over the decision trees.

Just like the forest models, the gradient boosting model should have improved predictive accuracy because of variance reduction. It is hoped that the final model will have low bias and low variance.



Neural Network Models

The neural network model is a natural extension of regression models. Neural network models are arranged in layers. The first layer, which is called in the input layer, consists of the predictor variables. The second layer, which is called the hidden layer, consists of the hidden units. The third layer is the target layer which consists of the response.



Neural Network Diagram

bias estimate

hidden unit

$$\hat{y} = \hat{w}_{00} + \hat{w}_{01} \cdot H_1 + \hat{w}_{02} \cdot H_2 + \hat{w}_{03} \cdot H_3$$

prediction estimate

weight estimate

hidden units

$$\begin{cases} H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} \cdot x_1 + \hat{w}_{12} \cdot x_2) \\ H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} \cdot x_1 + \hat{w}_{22} \cdot x_2) \\ H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} \cdot x_1 + \hat{w}_{32} \cdot x_2) \end{cases}$$

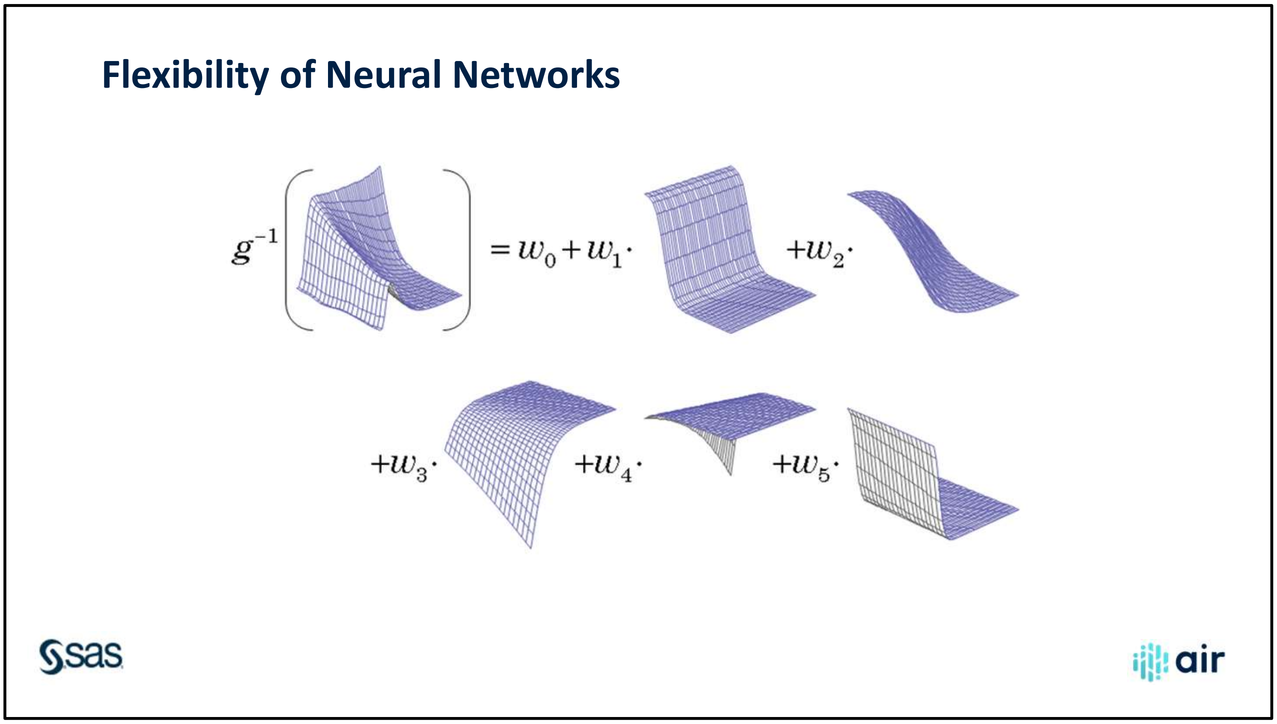
activation function

The prediction equation consists of the bias estimate and the weight estimates for each of the hidden units.

The hidden units can be thought of as regressions on the original predictor variables that the neural network figures out. The hidden units include a link function, which is called an activation function in neural networks. In the example, the activation function is the hyperbolic tangent function which rescales the output from the hidden units to be between -1 and 1.

The weight estimates are computed using least squares estimation for interval targets and maximum likelihood for categorical targets.

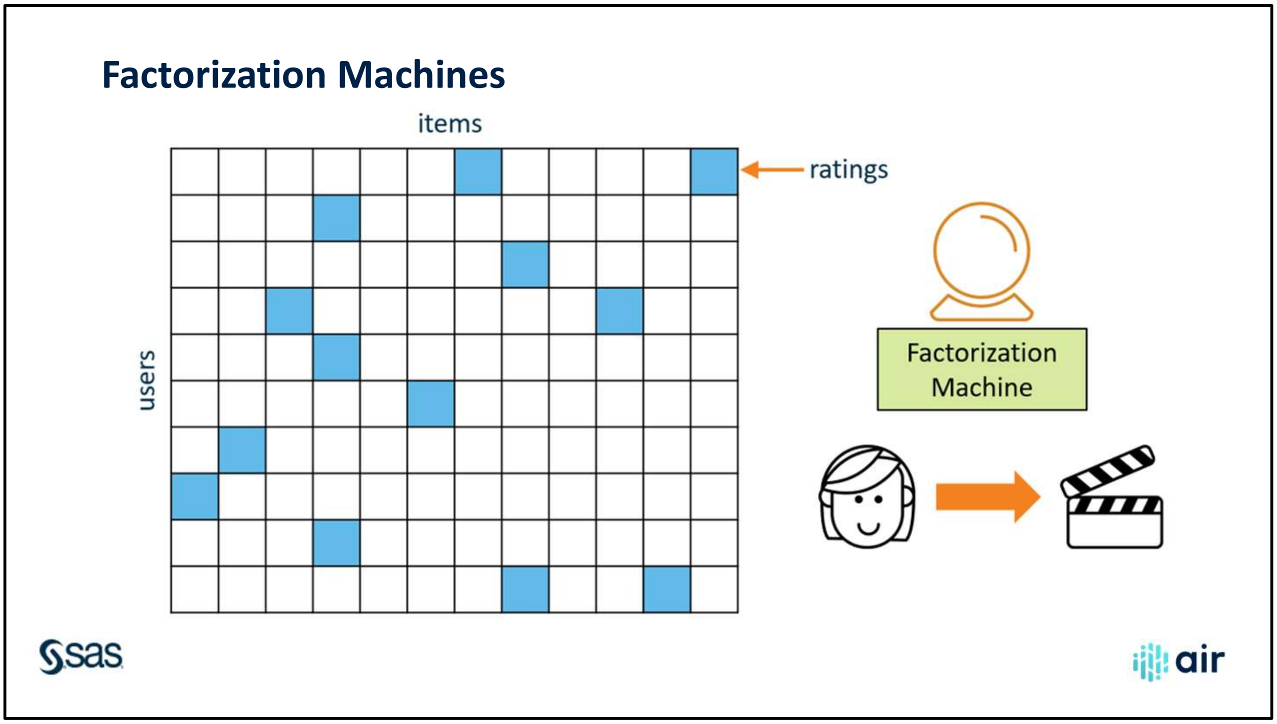
After the prediction formula is generated, obtaining a prediction is simply a matter of plugging the predictor variable values into the hidden unit expressions. In the same way as regression models, data scientists obtain the prediction estimates using the appropriate link function in the prediction equation.



Flexibility of Neural Networks

The chief benefit of neural networks is their unlimited flexibility. A neural network is a universal approximator, which means that with enough hidden units, the neural network can model any predictor variable and target relationship no matter how complex.

The output from a neural network with one hidden layer is a weighted linear combination of the mathematical functions generated by the hidden units. The weights and biases give these functions their flexibility. Changing the orientation and steepness of these functions, and then combining them, enables the neural network to fit any target.



Factorization Machines

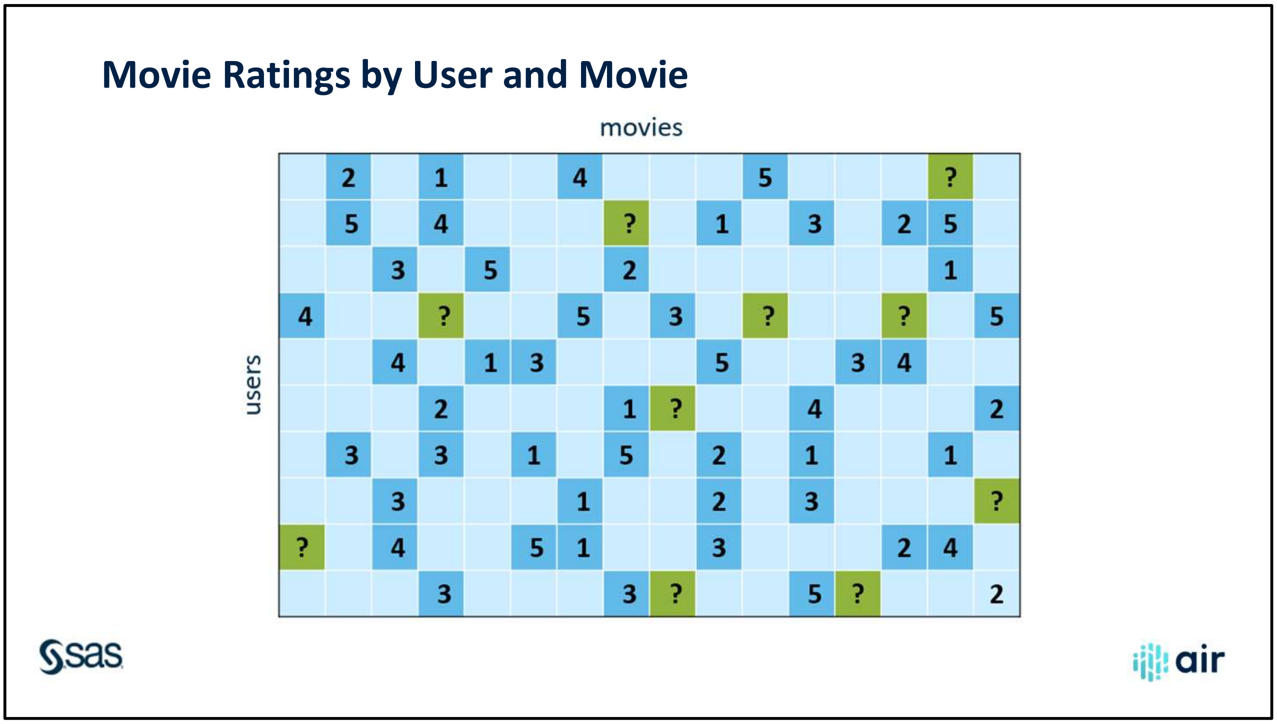
items

ratings

users

Factorization Machine

Factorization machines are useful when data scientists have a huge matrix with sparse data. For example, a company might want to learn about user preferences in order to recommend items such as movies, books, or songs. The purpose is to predict which ratings a user would give to a set of items. In this business scenario, companies do not have much information about the users or about the items. The main goal is to evaluate possible relationships between users and items.



Movie Ratings by User and Movie

movies

users

	2		1			4				5				?	
	5		4				?		1		3		2	5	
		3		5			2							1	
4			?			5		3		?			?		5
		4		1	3				5			3	4		
			2				1	?			4				2
	3		3		1		5		2		1			1	
		3				1			2		3				?
?		4			5	1			3				2	4	
			3				3	?			5	?			2

Consider a web-based company that sells on-demand movies to users. This hypothetical company basically has no information about its customers (users) except the movies (items) that they download and watch, and eventually, their ratings for each movie.

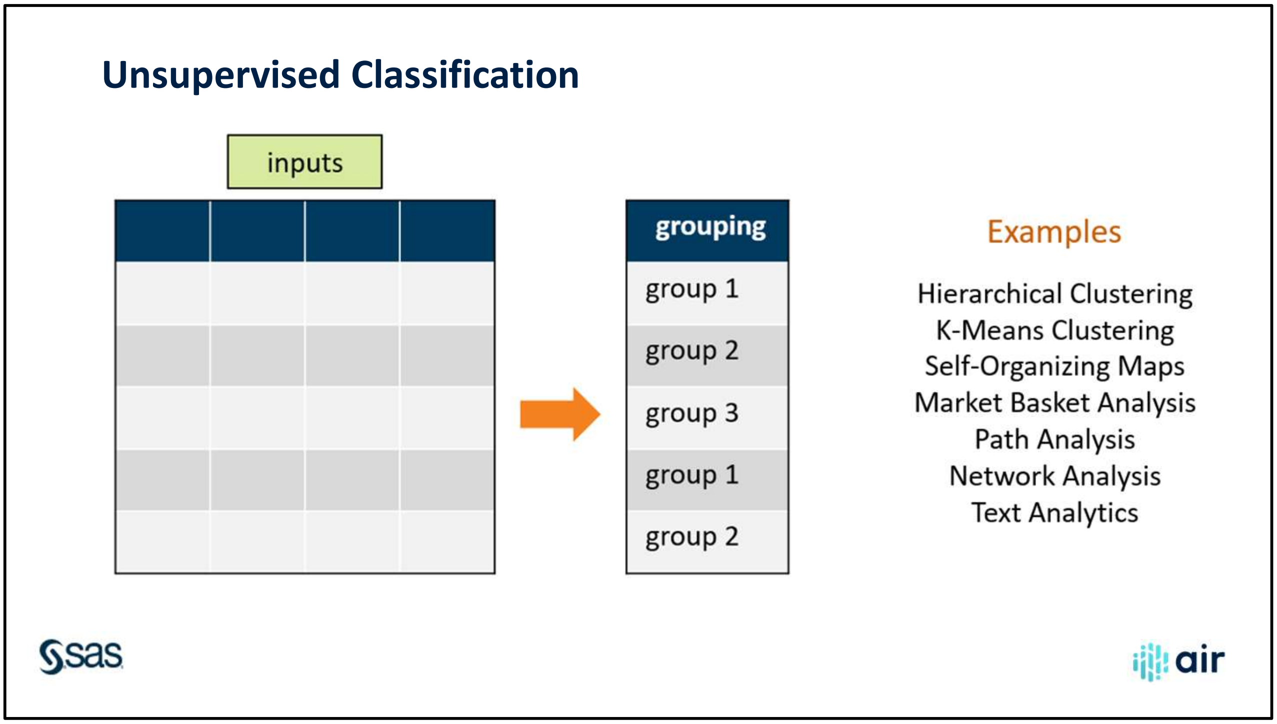
In this example, data scientists can create a huge matrix that relates all users to all movies. As expected, most of the cells in that matrix would be missing. Not all users give ratings to all the movies that they watch. The challenge is to estimate a user's rating for a movie. The intuition factorization machines use to solve this problem is that there should be some latent features that determine how a user rates a movie.

Factorization machine models that are used in recommender systems where the aim is to predict user ratings on items. Assume there are two categorical variables: **u** for users and **i** for movies. The input vector is constructed using binary indicator variables for the user and item.

The dot product of the vectors for the users and items models the interaction between the two variables. However, instead of using an interaction parameter, the factorization machine models the interaction using matrix factorization.
In sparse settings, there is not enough data to estimate the interactions between the users and the items. Factorization machine models get around this problem by breaking the

independence of the interaction parameters by factorizing them. In general, this means that the data for one interaction helps to estimate the parameters for related interactions. This is accomplished using matrix factorization techniques, which enables the model to discover the latent features underlying the interactions between users and items. The basic idea is to discover two matrices that, when multiplied together, returns the original matrix.

It should be noted that factorization machine models are predictive models like support vector machines and are not restricted to just recommender systems. This more general model would look different from what is presented here.



Unsupervised Classification

inputs

grouping
group 1
group 2
group 3
group 1
group 2

Examples

Hierarchical Clustering
K-Means Clustering
Self-Organizing Maps
Market Basket Analysis
Path Analysis
Network Analysis
Text Analytics

Unsupervised classification is a collection of methods that do not have a target variable and try to find previously unknown patterns in a data set. An algorithm used in these methods is clustering, which attempts to group cases in the data based on the similarities of the input variables.

There are many successful applications of unsupervised models:

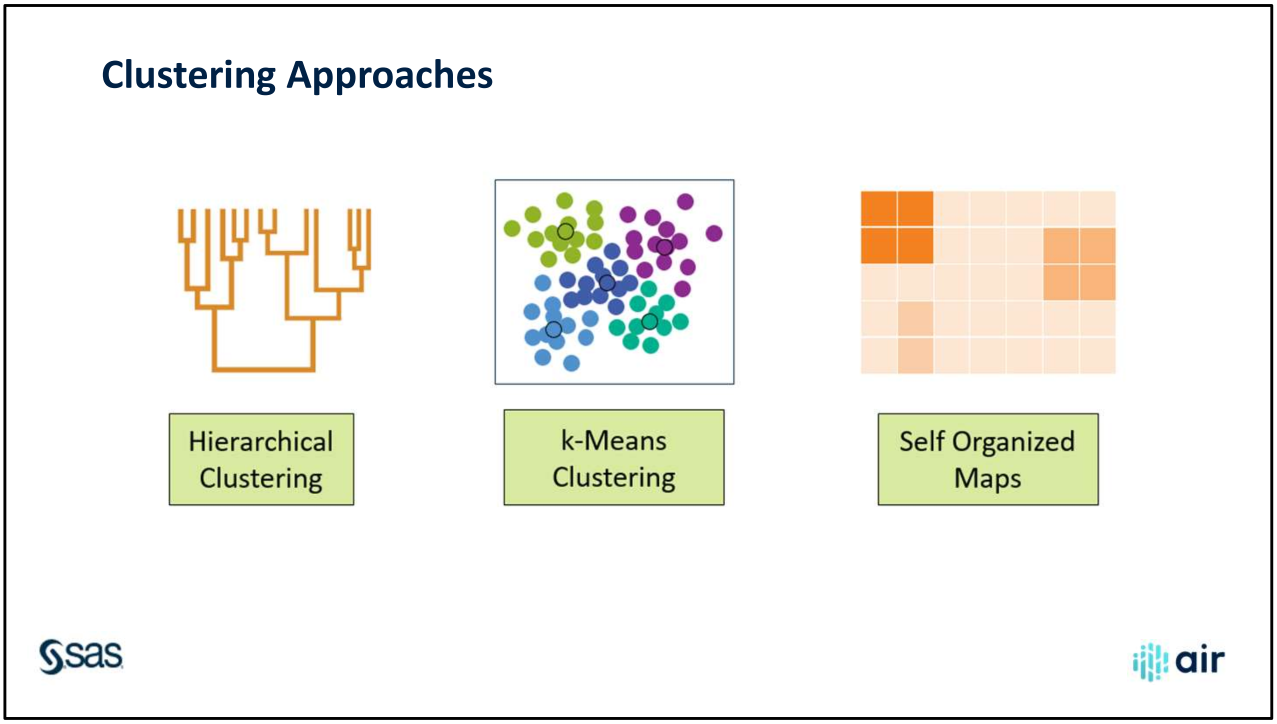
• Data reduction exploits patterns in the data to create a more compact representation of the original. Though vastly broader in scope, data reduction includes analytic methods such as cluster analysis.

• Novelty detection methods, also known as *anomaly detection*, seek unique or previously unobserved data patterns. The methods find application in business, science, and engineering. Business applications include fraud detection, warranty claims analysis, and general business process monitoring.

• Profiling is a by-product of reduction methods such as cluster analysis. The idea is to create rules that isolate clusters or segments, often based on demographic or behavioral measurements. Data scientists might develop profiles of a customer database to describe the consumers of a company’s products.

- Market basket analysis, or *association rule discovery*, is used to analyze streams of transaction data for *combinations* of items that occur (or do not occur) more (or less) commonly than expected. Data scientists can use this to identify interesting combinations of purchases or as predictors of customer segments.

- Sequence analysis is an extension of market basket analysis to include a time dimension to the analysis. In this way, transactions data are examined for *sequences* of items that occur (or do not occur) more (or less) commonly than expected. Data scientists might use sequence analysis to identify patterns or problems of navigation through a website. It should be noted that sequence analysis is different from time series analysis and natural language processing. The data in those analyses are a sequence of text strings or numbers. The data in sequence analysis are short sequences of customer behavior.



Clustering Approaches

Hierarchical Clustering

k-Means Clustering

Self Organized Maps

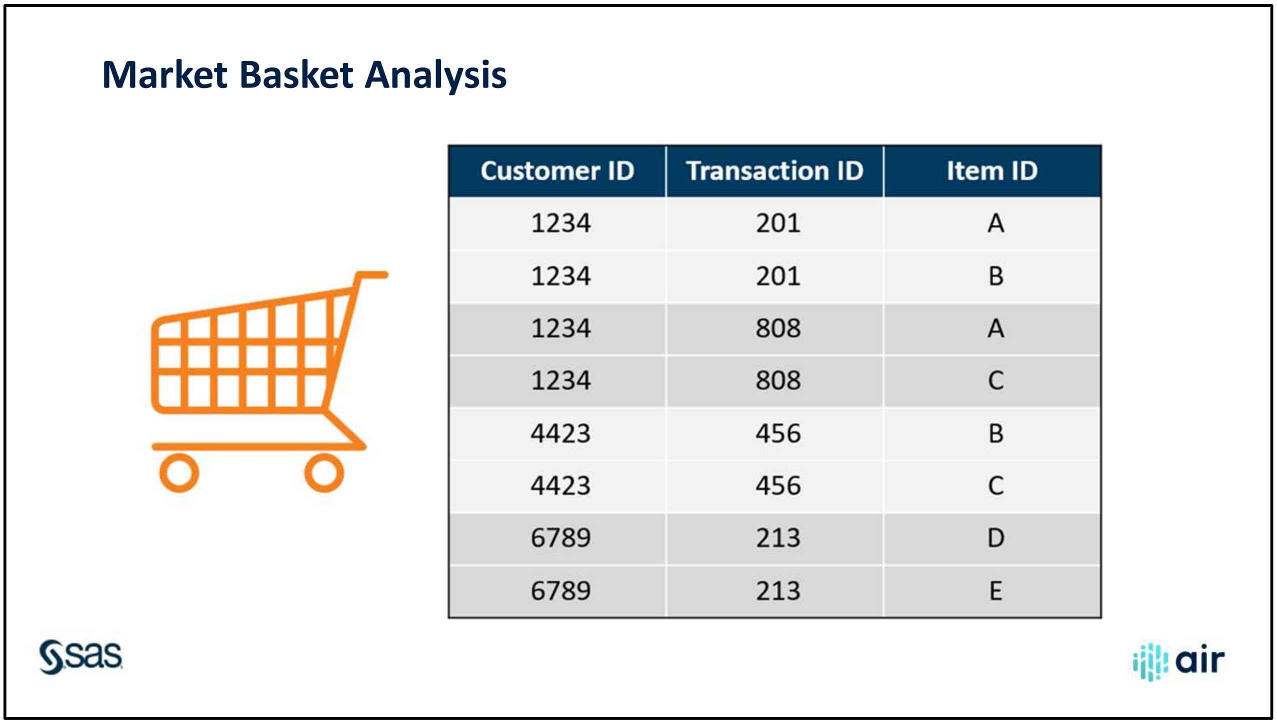
The goal of clustering is to partition data into groups so that the observations within a group are as similar as possible to each other, and as dissimilar as possible to the observations in other groups. It is a data reduction method because an entire training data set can be represented by a small number of clusters. The groupings are known as clusters or segments, and they can be applied to other data sets to classify new cases.

Examples of clustering methods are:

Hierarchical clustering creates clusters that are hierarchically nested within clusters at earlier iterations. Agglomerative clustering starts with one cluster per point, and repeatedly merges nearby clusters.

k-Means clustering divides a data set into clusters by trying to minimize some specified error function. K-means clustering starts with k clusters and assigns the data points to the nearest center. The algorithm shifts centroids and points over time until no more moves are needed.

Self organized maps assign observations to clusters based on the similarities of their attributes. Self-organizing maps are neural networks that provide a topological mapping from the input space to the clusters. Every observation assigned affects the cluster.



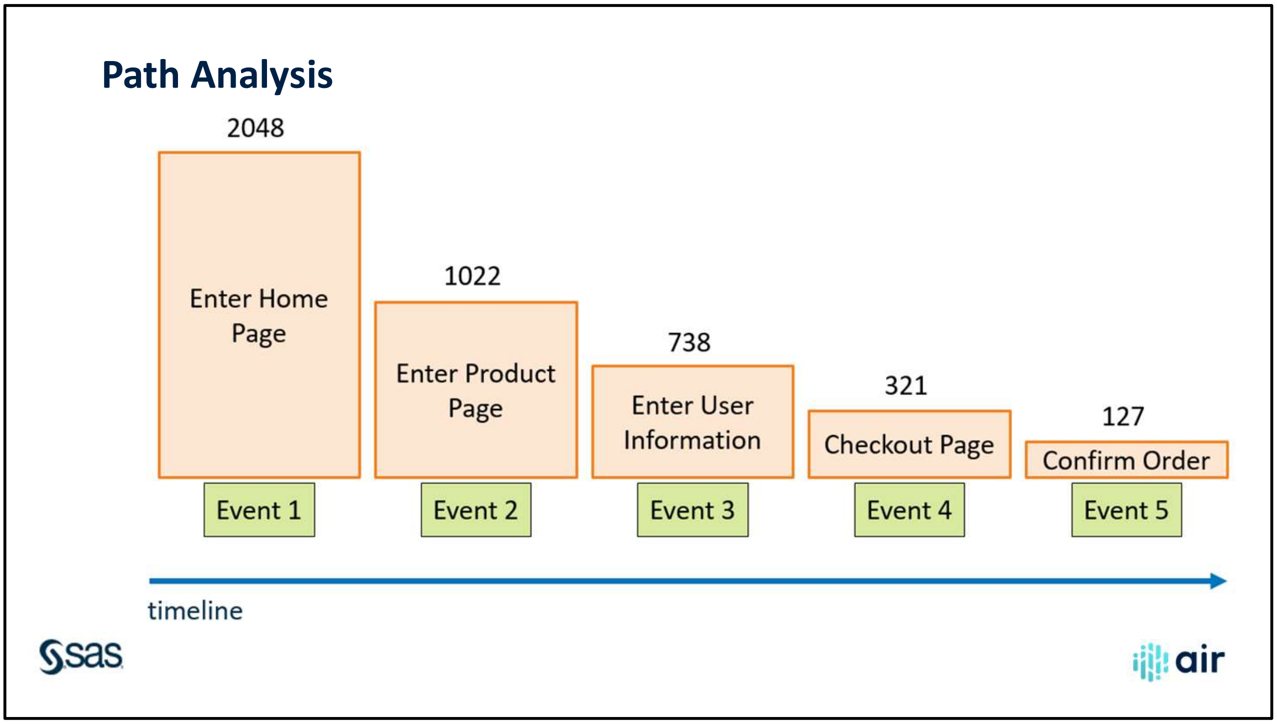
Market Basket Analysis

Customer ID	Transaction ID	Item ID
1234	201	A
1234	201	B
1234	808	A
1234	808	C
4423	456	B
4423	456	C
6789	213	D
6789	213	E

Market basket analysis, or association rule discovery, is used to analyze streams of transaction data for combinations of items that occur (or do not occur) more (or less) commonly than expected. Data scientists can use market basket analysis to identify interesting combinations of purchases, or as predictors of customer segments.

In the simplest situation, the data consists of two variables: a transaction and an item. For each transaction, there is a list of items. Typically, a transaction is a single customer purchase, and the items are the things that were purchased. An association rule is a statement of the form (item set A) => (item set B). In other words, an association rule is a pattern that states when A occurs, B occurs with a specified probability.

The aim of the analysis is to determine the strength of all the association rules among a set of items.



Path Analysis

2048
Enter Home Page
Event 1

1022
Enter Product Page
Event 2

738
Enter User Information
Event 3

321
Checkout Page
Event 4

127
Confirm Order
Event 5

timeline

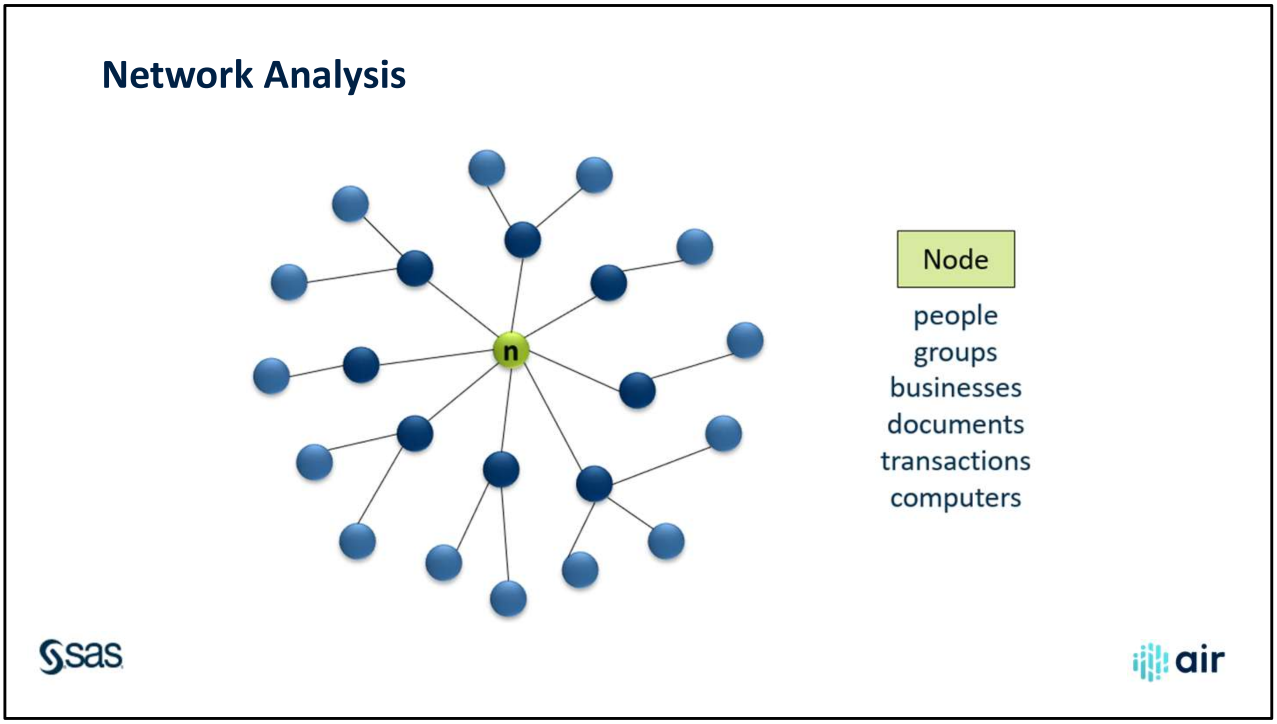
Path analysis is a portrayal of a chain of consecutive events that a given user or cohort performs during a set period. It is a way to understand user behavior in order to gain actionable insights into the data. For example, if a web site offers products for sale, the owner wants to convert as many visitors to a completed purchase as possible. Data scientists can use path analysis to determine what features of the website encourage the desired result. They can also look for “black holes,” which are paths or features that lose or confuse potential customers.

It is important to note that the path analysis referred to here is not the same as the path analysis that is a special case of structural equation modeling.

Data scientists can use path analysis to examine the customer experience with online games (for example, if a large amount of time is spent wandering around the menu page, there might be a problem with the user experience). By following the customer path, data scientists can detect problems and fix the error.

Data scientists can use path analysis to customize an e-commerce shopping experience. When presented with products other customers in a certain cohort looked at, a platform can suggest “items you might like”.
Finally, path analysis can help solve performance issues on a server. Data scientists can

examine a path and realize that its site freezes up after a certain string of events. The error can then be documented and fixed.



Network Analysis

Node

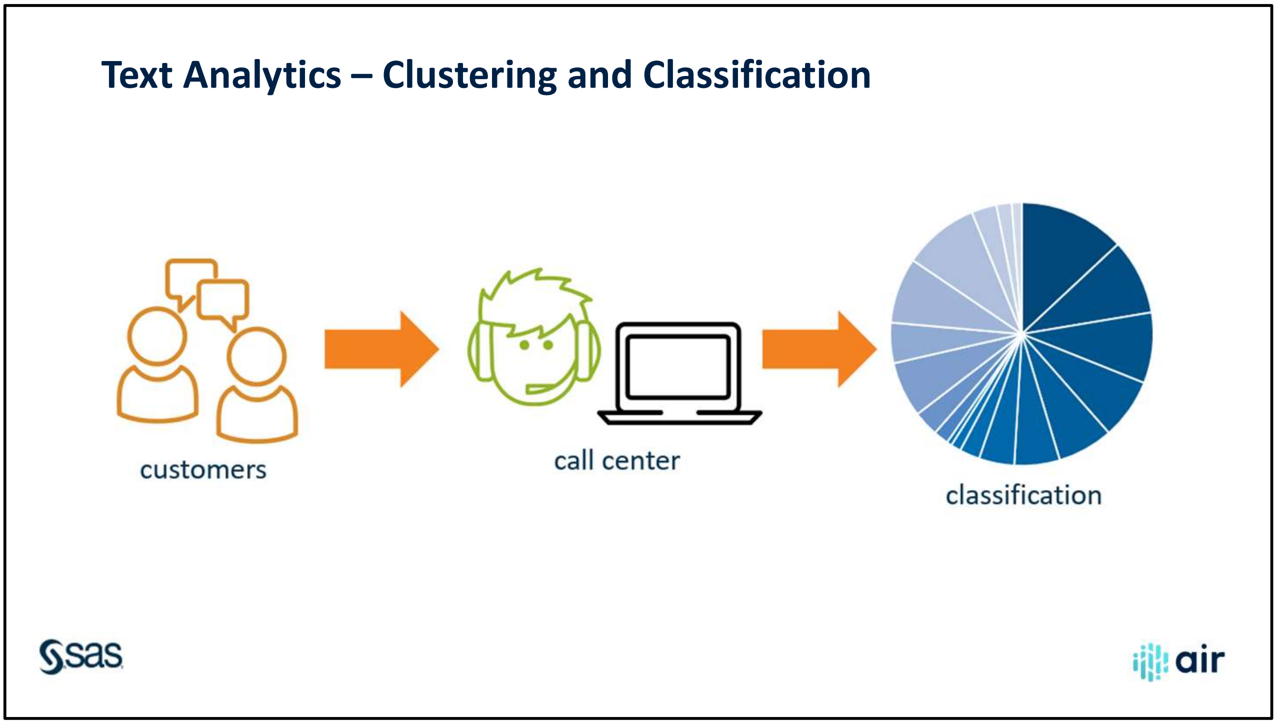
people
groups
businesses
documents
transactions
computers

Network analysis is the mapping and measuring of relationships and flows between nodes. The nodes in the network can be people, groups, businesses, documents, transactions, computers, and so on. The links show relationships or flows between the nodes. Network analysis enables us to identify and understand the importance of certain nodes within a network, as well as identify clusters and patterns within networks.

Companies use network analysis to address relevant business events such as churn, fraud, facility location, in addition to strategies related to increasing sales, product adoption, and service consumption.

Governments use network analysis to tackle terrorism, money laundering, and fraud in tax payment or in health care, among many other applications.

A common example of network analysis is social network analysis, which is the process of investigating social structures. Examples of social structures commonly visualized through social network analysis are social media networks, information circulation, friendship networks, and disease transmission (for example, a node can be a person and a link is whether two people are friends).

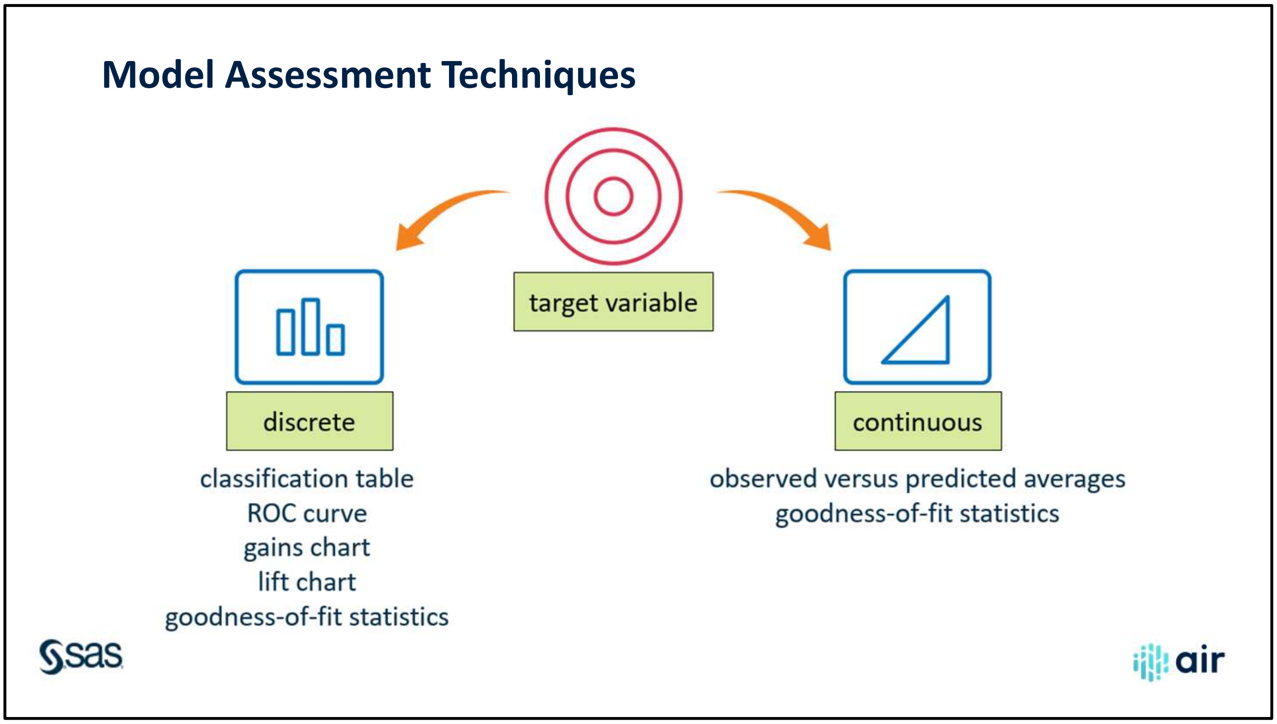


Text Analytics – Clustering and Classification

Data scientists use text analytics to analyze unstructured text, extract relevant information, and transform it into useful business intelligence. Text analytics can be used to create a set of predictor variables that contain information about what is written in the text. It can also be used to cluster the topics that are mentioned in the documents.

Call center optimization based on the types and volumes of calls is a business problem that can be addressed by text analytics. The customers contact the call center for a variety of reasons dealing with topics. The call center representative requires subject matter knowledge to address these different topics. Data scientists can use text analytics to cluster the topic areas and to estimate the optimal number of call center representatives for each topic.

With the use of text analytics, data scientists can cluster the various reasons customers call the call center. This information can be used to train the call center representatives to deal with topics such as broadband failure, chip failure, internet failure, and so on. The result will be an adequate number of call center representatives to address the customer requests.



Model Assessment Techniques

target variable

discrete

classification table
ROC curve
gains chart
lift chart
goodness-of-fit statistics

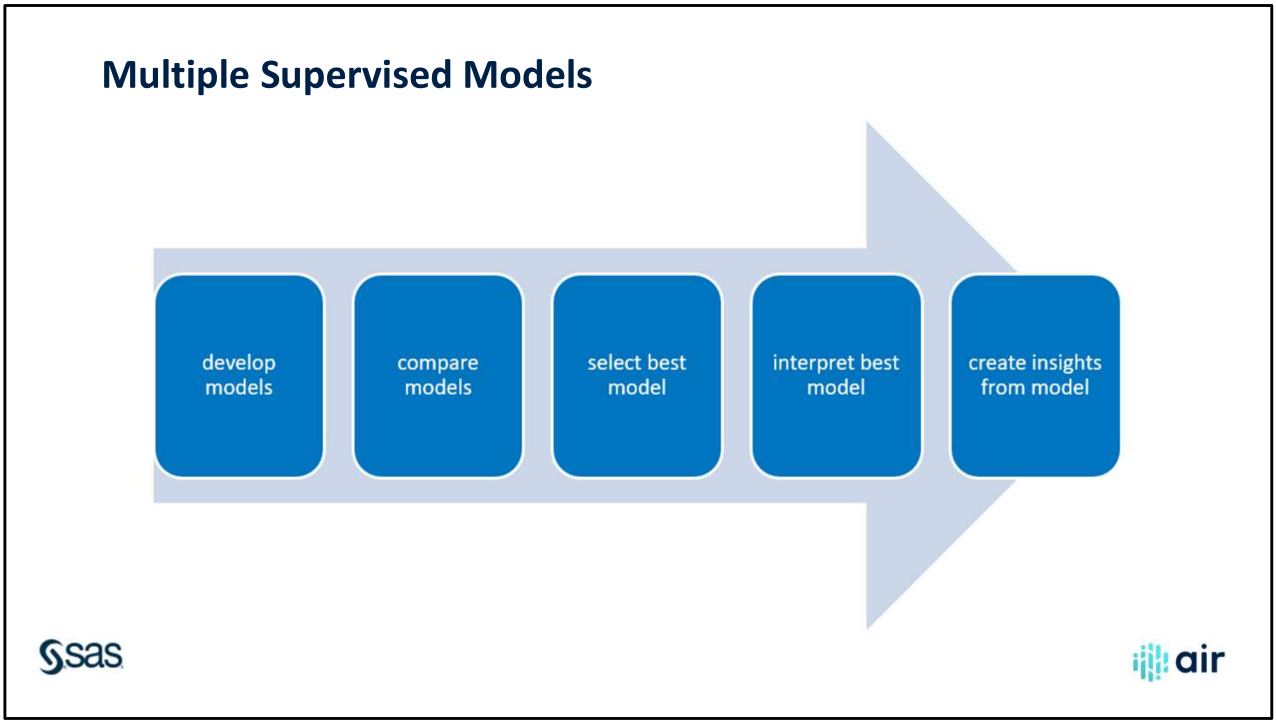
continuous

observed versus predicted averages
goodness-of-fit statistics

The model assessment techniques depend on the scale of the target variable. If the target variable is discrete, then classification tables, ROC curves, gains and lift charts, and several goodness-of-fit statistics can be used to assess model performance. Some of the goodness-of-fit statistics include the misclassification error rate, information criteria such as the AIC, and the Brier score which is the weighted squared difference between the predicted probabilities and their observed response levels.

If the target variable is continuous, a plot of the observed versus predicted averages by decile would be useful. Goodness-of-fit statistics such as the average squared error and the adjusted R-square could also be used.

For categorical targets, cases are allocated to classes based on cutoff values of the predicted probability. The steps include the following:
1. Estimate the predicted probability of class 1 for each case by the logistic regression model.
2. Choose a cutoff probability.
3. Assign cases to class 1 if their estimated predicted probability exceeds the cutoff. Otherwise, assign the case to class 0.



Multiple Supervised Models

develop models

compare models

select best model

interpret best model

create insights from model

A normal approach in data science is to try out multiple models, considering different methods and distinct algorithms. The steps are as follows:

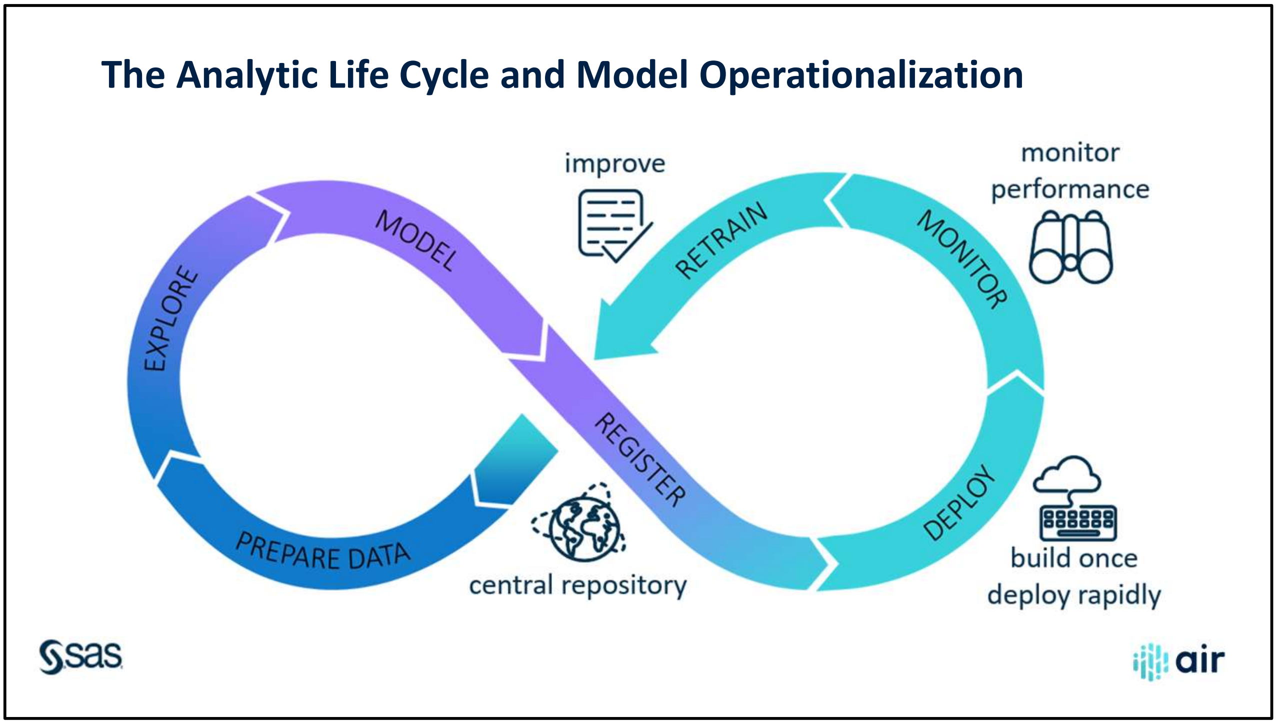
1. After all models are trained, they need to be evaluated according to the business needs and the possible deployment actions. All these facts will lead to the champion model.

2. All trained models can be compared and evaluated based on multiple fit statistics, depending on the business requirements.

3. The best model can be selected based on its performance on test and validate data in order to generalize its performance in production.

4. Model interpretability plots or decision trees can be used to interpret black-box models.

5. Insights about all the models and correlations between input data and the target can lead to more accurate business actions.



The Analytic Life Cycle and Model Operationalization

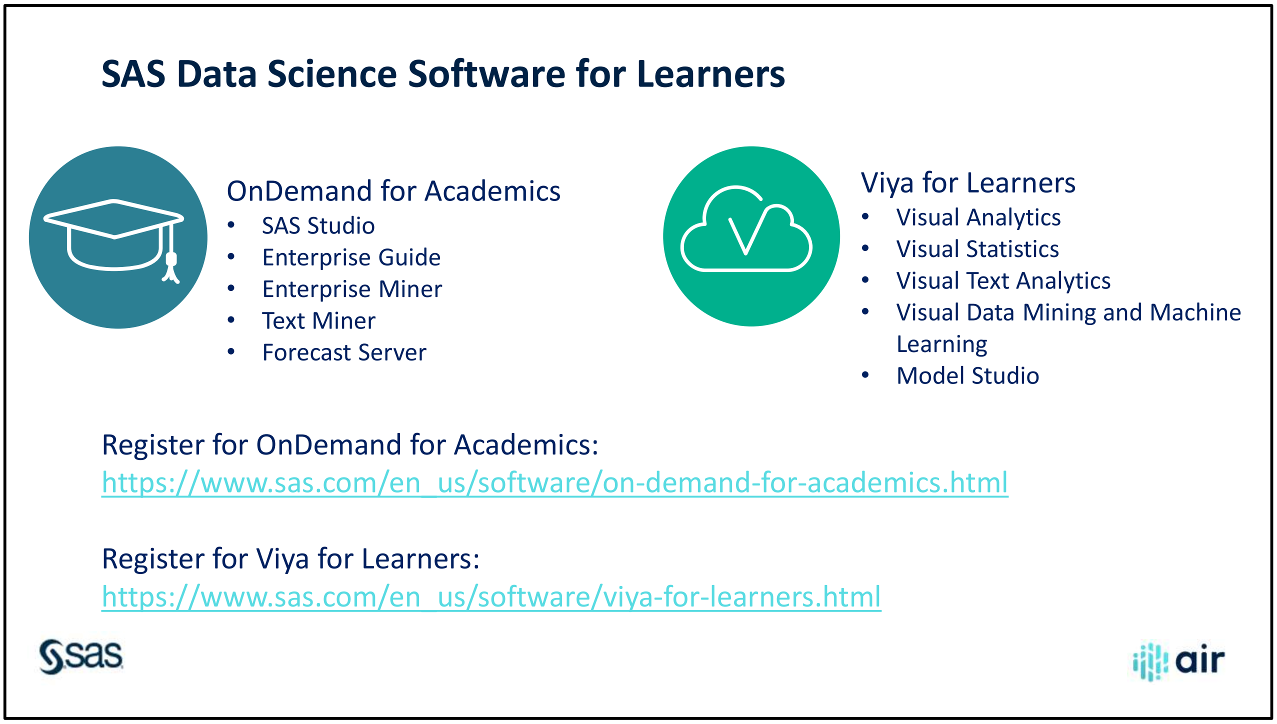
The analytical life cycle now involves model operationalization. When the champion models from each distinct modeling approach are compared and evaluated, these are the next steps:

1. Champion and challenger models should be registered and published for future deployment. Registering models means putting the models in a central repository. Publishing the model means putting the score code where the model is run in production.

2. Models should be deployed by score codes and run on multiple platforms. Models need to score data in production considering new data over time.

3. Models should be monitored to evaluate their performance over time. Models should be easily managed in an enterprise and collaborative manner.

4. If the model performance decreases over time, then the models need to be retrained with new data and the analytical cycle starts all over again.



SAS Data Science Software for Learners

OnDemand for Academics

- SAS Studio
- Enterprise Guide
- Enterprise Miner
- Text Miner
- Forecast Server

Viya for Learners

- Visual Analytics
- Visual Statistics
- Visual Text Analytics
- Visual Data Mining and Machine Learning
- Model Studio

Register for OnDemand for Academics:
https://www.sas.com/en_us/software/on-demand-for-academics.html

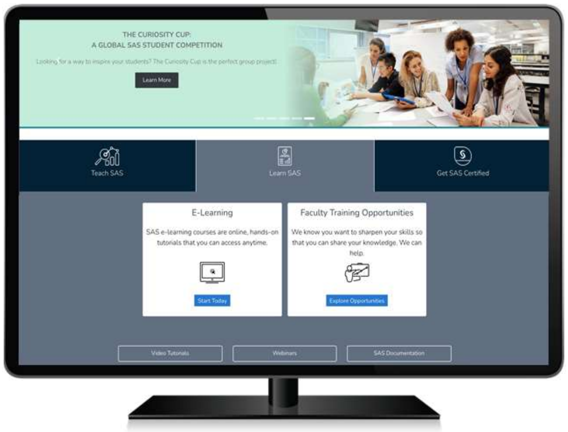
Register for Viya for Learners:
https://www.sas.com/en_us/software/viya-for-learners.html

Register for OnDemand for Academics:
https://www.sas.com/en_us/software/on-demand-for-academics.html

Register for Viya for Learners:
https://www.sas.com/en_us/software/viya-for-learners.html

SAS Educator Portal

- Instructional materials for data science, machine learning, statistics, and more!
    - 100+ Teaching kits
    - Industry-specific classroom activities
- E-learning & Certification Paths
- Educator Workshops & Training
- Discounts & Promotions



THE CURIOSITY CUP:
A GLOBAL SAS STUDENT COMPETITION

Looking for a way to inspire your students? The Curiosity Cup is the perfect group project!

E-Learning

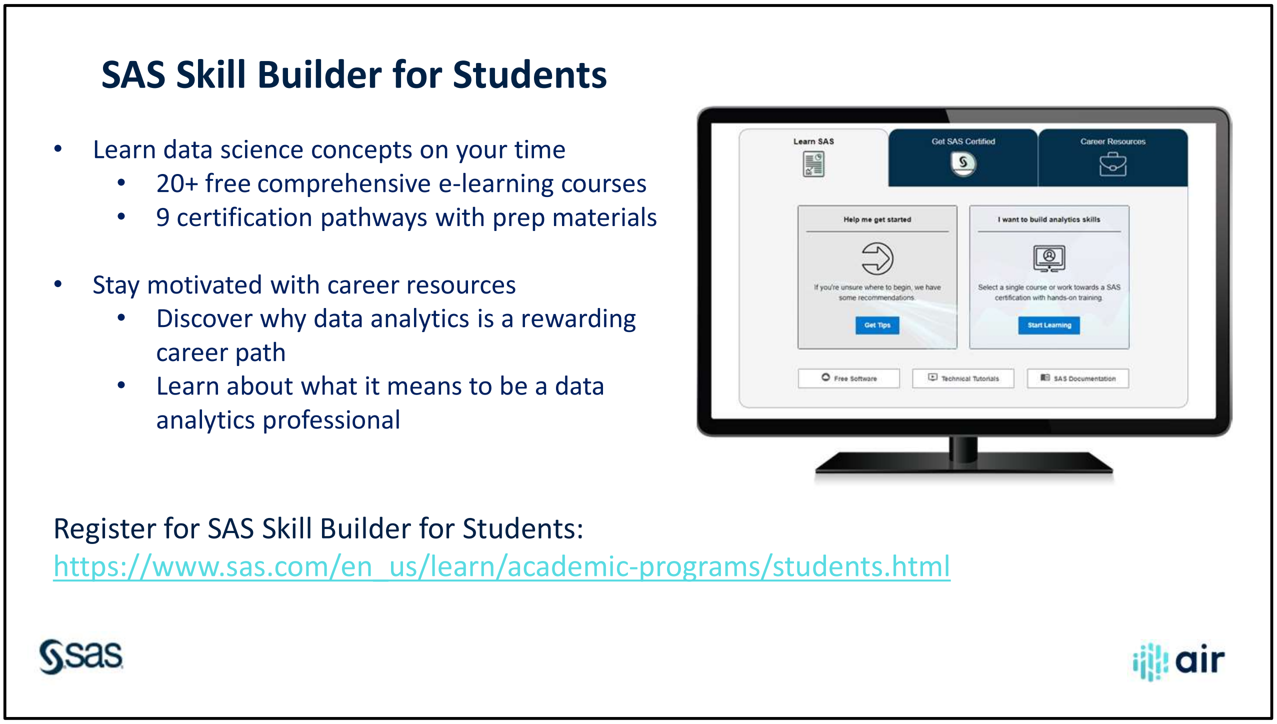
SAS e-learning courses are online, hands-on tutorials that you can access anytime.

Faculty Training Opportunities

We know you want to sharpen your skills so that you can share your knowledge. We can help.

Register for the SAS Educator Portal:
https://www.sas.com/en_us/learn/academic-programs/educators.html

Register for the SAS Educator Portal:
https://www.sas.com/en_us/learn/academic-programs/educators.html



SAS Skill Builder for Students

- Learn data science concepts on your time
  - 20+ free comprehensive e-learning courses
  - 9 certification pathways with prep materials
- Stay motivated with career resources
  - Discover why data analytics is a rewarding career path
  - Learn about what it means to be a data analytics professional

If you're unsure where to begin, we have some recommendations.

Select a single course or work towards a SAS certification with hands-on training.

Register for SAS Skill Builder for Students:
https://www.sas.com/en_us/learn/academic-programs/students.html

Register for SAS Skill Builder for Students:
https://www.sas.com/en_us/learn/academic-programs/students.html

Thank you!

Questions?
Jacqueline.Johnson@sas.com

Jacqueline Johnson, SAS Global Academic Programs
Jacqueline.Johnson@sas.com