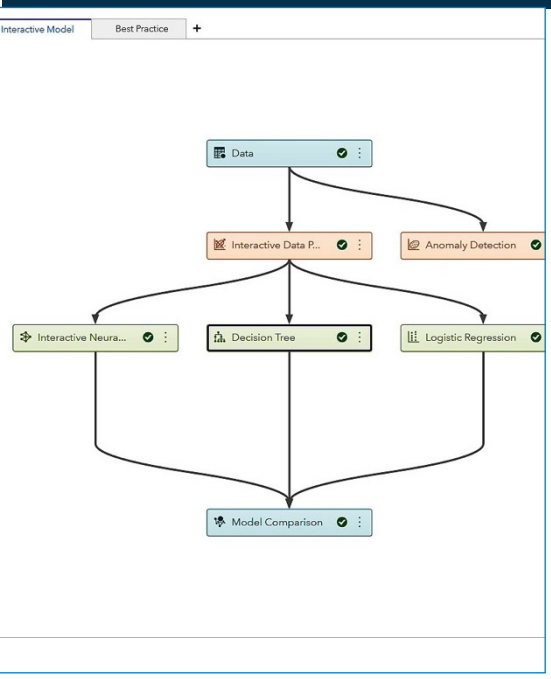
SAS® Visual Data Mining and Machine Learning

Все, что необходимо для решения самых сложных аналитических задач — в едином интегрированном решении для совместной работы.



Какие преимущества дает решение SAS® Visual Data Mining and Machine Learning?

Данное решение предоставляет полнофункциональный графический интерфейс, который обеспечивает выполнение всех этапов анализа данных. Помимо внедрения инновационных технологий машинного обучения для анализа структурированных и неструктурированных данных, решение позволяет включить в аналитический процесс и все остальные задачи. Специалисты получают возможность работать в единой интегрированной среде на всех этапах — от подготовки и изучения данных до разработки и развертывания модели. Масштабируемые и гибкие инструменты обработки позволяют ускорить решение самых сложных задач.

Что делает решение SAS® Visual Data Mining and Machine Learning незаменимым?

SAS Visual Data Mining and Machine Learning — первое решение, в котором инструменты продвинутой аналитики, подготовки данных, визуализации, оценки и развертывания модели объединены в рамках единой среды. Решение также поддерживает программирование на самых распространенных языках с открытым исходным кодом. Общая среда для совместной работы позволяет получать воспроизводимые результаты, что полезно для усовершенствования организационных процессов и выявления новых возможностей для роста и развития.

На кого ориентировано решение SAS® Visual Data Mining and Machine Learning?

Данное решение предназначено для всех, кто анализирует большие многоуровневые массивы данных и создает предиктивные модели. К таким специалистам относятся эксперты в области анализа и обработки данных, статистики, специалисты по интеллектуальному анализу данных, бизнес-аналитики, инженеры данных и исследователи.

Масштабы собранных данных неуклонно растут. Наблюдается нехватка высококлассных специалистов по анализу и обработке данных. Компаниям приходится в сжатые сроки решать все более и более сложные проблемы. Неважно, идет речь об анализе каждой транзакции для выявления мошеннических действий, или об анализе постоянно растущих объемов информации в социальных сетях для улучшения качества обслуживания клиентов, или о создании быстрой и точной системы рекомендаций для прогнозирования оптимальных предложений для клиентов — многофункциональное программное обеспечение на основе технологий машинного обучения поможет решить все эти задачи.

Решение SAS Visual Data Mining and Machine Learning реализует все этапы преобразования необработанных данных в полезную информацию, используя интегрированный графический интерфейс. Благодаря этому решению аналитики смогут гораздо быстрее получить доступ к данным и подготовить их для анализа, сформировать признаки, выполнить разведочный анализ данных, создать и сравнить модели машинного обучения, а также создать скоринговый код для внедрения предиктивных моделей.

Преимущества

- **Повышение производительности команды аналитиков.** Поддерживая все последовательные этапы машинного обучения, данное решение позволяет пользователям создавать и развертывать сложные модели в единой среде для совместной работы, что гарантирует высокую точность результатов.
- **Сокращение времени между получением данных и принятием решений.** Интерактивные графические и программные интерфейсы позволяют значительно сократить время, затрачиваемое на подготовку данных и создание моделей. Высокая скорость обработки обеспечивает быстрые результаты.
- **Возможность изучить альтернативы при поиске оптимального решения.** Максимальная производительность, доступная за счет распределенных вычислений, и функциональные компоненты конвейера машинного обучения позволяют неограниченному числу пользователей быстро изучить и сравнить несколько альтернативных вариантов решения. Автоматическая настройка обеспечивает возможность тестирования различных сценариев для определения оптимальной модели. Воспроизводимость всех этапов анализа гарантирует полноту и достоверность результатов и выводов.
- **Сложные аналитические задачи решаются быстрее**. Данное решение создано на основе SAS® Viya® — новейшего дополнения к платформе SAS, которое обеспечивает максимально быстрое создание предиктивных моделей и внедрение методов машинного обучения. Данные размещаются в оперативной памяти, поэтому в ходе итеративного анализа их не придется загружать несколько раз. Обработка аналитических моделей занимает максимум несколько минут, а не несколько часов, как это было раньше, поэтому теперь вы сможете найти решения сложных задач гораздо быстрее.
- **Быстрое развертывание моделей при помощи автоматического скорингового кода SAS.** Временные затраты дополнительно снижаются благодаря простому в использовании скоринговому коду, который может быть вызван как из SAS, так и из Python, R, Lua, Java, а также через REST API.
- **Возможность использовать различные языки программирования.** Программисты на Python, R, Java и Lua смогут по достоинству оценить возможности данного решения без необходимости специально изучать SAS-программирование. Они получают доступ к надежным и проверенным алгоритмам машинного обучения SAS и смогут использовать их при программировании на других языках.

Обзор

SAS Visual Data Mining and Machine Learning предлагает новую мощную полнофункциональную графическую среду, которая обеспечивает поддержку всех возможностей машинного обучения и глубокого обучения — от доступа к данным и их обработки до построения и развертывания сложных моделей. Распределенные вычисления, выполняемые в оперативной памяти, позволяют обрабатывать большие объемы данных и строить сложные модели, быстро решать поставленные задачи и максимально эффективно использовать ресурсы.

Гибкая и удобная графическая среда для аналитики

В настоящее время неограниченное количество пользователей могут анализировать любые объемы структурированных и неструктурированных данных с помощью простого графического интерфейса. Каждый проект (и его конечная задача) определяется визуальными конвейерами; они позволяют разбить жизненный цикл анализа на серию логически упорядоченных этапов. Отдельные ветви конвейеров могут выполняться асинхронно. Графический интерфейс (Model Studio) предоставляет интегрированную среду для выполнения типовых операций машинного обучения: подготовка данных, формирование признаков, исследование данных, построение и развертывание моделей. Интерактивные интерфейсы для решения отдельных задач позволяют быстро и легко применять сложные алгоритмы к большим многоуровневым массивам данных. В результате этих операций создается код SAS, который можно сохранить для последующей автоматизации задач. Кроме того, можно предоставить общий доступ к фрагментам кода и шаблонам конвейеров. Model Studio предоставляет специализированную среду для совместной работы, которая поддерживает построение, развертывание и совместное использование моделей.

Масштабируемые инструменты аналитической обработки в оперативной памяти

Данное решение предоставляет безопасную многопользовательскую среду для одновременного доступа к данным в оперативной памяти. Операции с данными и аналитические запросы распределяются по различным узлам, где выполняется их многопоточная параллельная обработка. Это обеспечивает значительное повышение быстродействия. Все данные, таблицы и объекты хранятся в памяти в течение неограниченного срока, что обеспечивает максимально эффективную обработку. Встроенные функции обеспечения отказоустойчивости и управления распределением памяти позволяют запускать сложные процессы обработки данных и гарантируют их успешное выполнение.

Сокращается время аналитической обработки больших массивов данных, а также уменьшается объем сетевого трафика. Кроме того, доступны все преимущества современной многоядерной архитектуры, позволяющие быстрее найти решение поставленной задачи.

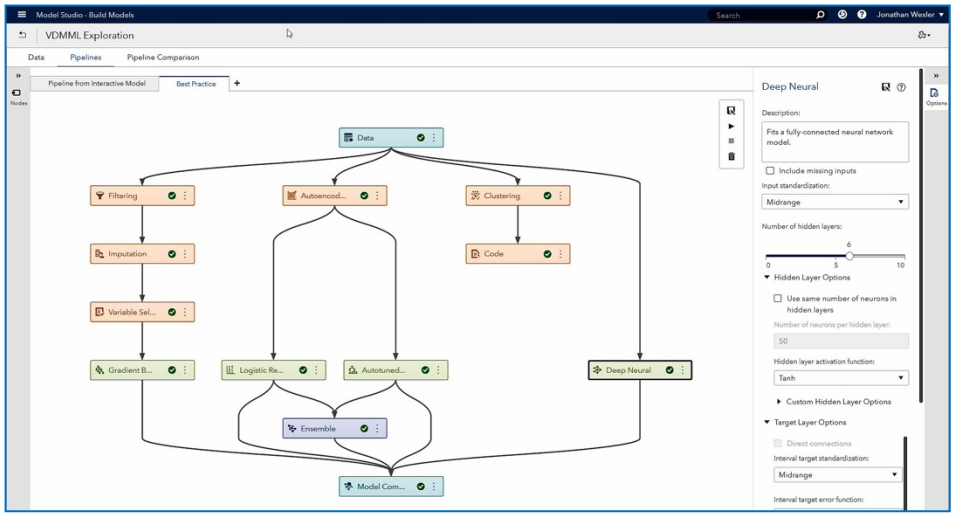
Ключевые функции

Интерактивное программирование в веб-среде разработки

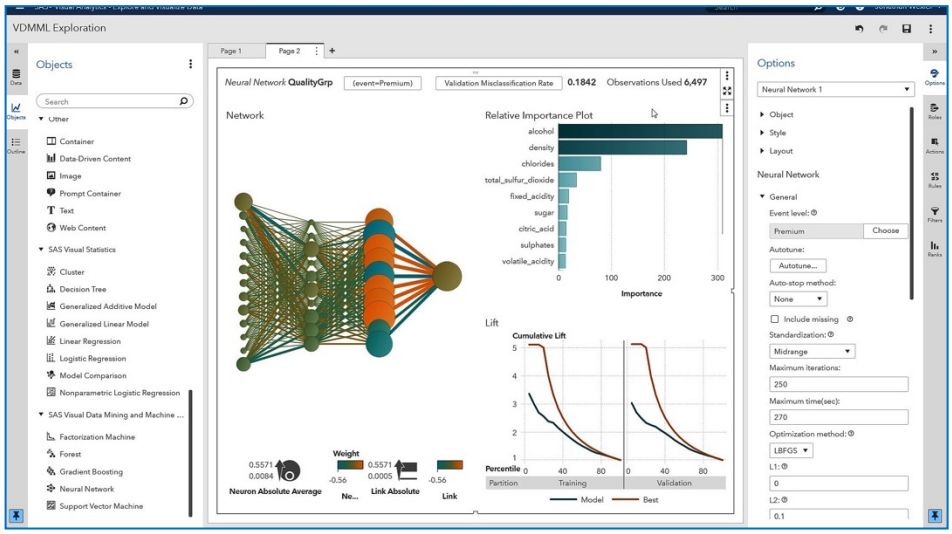
- Графический интерфейс для всех этапов анализа.
- Интерактивный интерфейс с поддержкой функции перетаскивания не требует использования кода, хотя такая возможность также поддерживается.
- Поддержка автоматического создания кода в каждом узле конвейера.
- Готовые шаблоны конвейеров различных уровней (базовый, средний и продвинутый) позволяют пользователям быстро приступить к решению задач машинного обучения.
- Среда для совместной работы обеспечивает возможность общего доступа участников проекта к данным, фрагментам кода и шаблонам конвейеров.

Масштабируемые инструменты для распределенных вычислений в оперативной памяти

- Распределенная обработка сложных аналитических расчетов для больших массивов данных в оперативной памяти обеспечивает низкую задержку по времени отклика.
- Аналитические запросы объединены в одну задачу, выполняемую в оперативной памяти, причём не требуется повторная загрузка данных или запись промежуточных результатов на диск.
- Одновременный доступ нескольких пользователей к данным в оперативной памяти существенно повышает эффективность работы.
- Данные и промежуточные результаты сохраняются в оперативной памяти в течение неограниченного срока, что уменьшает задержку по времени отклика.
- Встроенные инструменты управления рабочей нагрузкой обеспечивают эффективное использование вычислительных ресурсов.
- Встроенные инструменты обеспечения отказоустойчивости гарантируют успешное выполнение всех задач.



Визуальное представление конвейеров создаёт эффективную среду для совместной работы, в которой можно создавать и развертывать сложные модели машинного и глубокого обучения.



Решение SAS Visual Data Mining and Machine Learning дает пользователям возможность быстро разрабатывать и интерпретировать продвинутые алгоритмы машинного обучения.

Инновационные методы статистики, интеллектуального анализа данных и машинного обучения

Решение SAS Visual Data Mining and Machine Learning предоставляет расширенный набор современных статистических алгоритмов, а также алгоритмов машинного обучения, глубокого обучения и анализа текста в единой среде.

В аналитический инструментарий входят методы кластеризации, различные виды регрессии, модели случайного леса и градиентного бустинга, машины опорных векторов, методы обработки естественных языков, выделения тем из текстов и многое другое.
Все эти методы позволяют эффективно выявлять новые закономерности, тренды и взаимосвязи между данными в структурированных и неструктурированных массивах. Решение также предоставляет алгоритмы разложения матриц для построения индивидуально настраиваемых систем рекомендаций.

Благодаря возможности обработки больших объемов данных с высокой скоростью решение SAS Visual Data Mining and Machine Learning становится идеальным выбором для применения методов глубокого обучения. Среди таких алгоритмов – глубокие нейронные сети, сверточные нейронные сети для классификации изображений и рекуррентные нейронные сети, обеспечивающие усовершенствованный анализ текста.

Сложные алгоритмы обучения (например, нейронные сети, градиентный бустинг и случайный лес) можно автоматически настроить для достижения оптимального уровня эффективности, что позволяет значительно сэкономить время и ресурсы.

Интегрированные инструменты подготовки данных, исследования и формирования признаков

Для эффективного выполнения действий по подготовке аналитических данных, которые обычно занимают много времени, предусмотрен графический интерфейс с поддержкой функции перетаскивания, с помощью которого инженеры данных могут быстро выполнять преобразования, обогащать и объединять данные в рамках интегрированного визуального конвейера действий. Решение позволяет выявлять проблемы с данными и устранять их с помощью передовых аналитических методов. Также доступно выявление потенциальных предикторов, уменьшение размерности больших наборов данных и удобное конструирование новых признаков на основе исходных данных.

Ключевые функции (продолжение)

Разработка моделей с использованием современных алгоритмов машинного обучения

- Случайный лес:
  - Автоматизированное комбинирование деревьев решений для прогнозирования единой целевой переменной.
  - Автоматическое распределение независимых процессов обучения.
  - Поддержка «умной» автоматической настройки параметров модели.
  - Автоматическое создание скорингового кода на языке SAS для применения модели.
- Градиентный бустинг:
  - Автоматический итерационный поиск оптимального разбиения данных в зависимости от выбранной целевой переменной.
  - Автоматическое многократное формирование выборки входных данных с корректировкой весов по остаткам.
  - Автоматическое генерирование средневзвешенного значения для итоговой модели.
  - Поддержка бинарных, номинальных и интервальных целевых переменных.
  - Возможность настройки процедуры обучения деревьев решений с выбором количества деревьев, критерия разбиения узлов деревьев, глубины поддеревьев и вычислительных ресурсов.
  - Автоматические критерии остановки обучения на основе оценки качества модели по валидационной выборке для предотвращения переобучения.
  - Автоматическое создание скорингового кода на языке SAS для применения модели.
- Нейронные сети:
  - Автоматическая интеллектуальная настройка набора параметров для построения оптимальной модели.
  - Поддержка моделирования дискретных данных.
  - Нетривиальные значения параметров по умолчанию для большинства параметров нейронной сети.
  - Возможность выбора архитектуры нейронных сетей и весов.
  - Возможность использования произвольного количества скрытых слоев для использования глубокого обучения.
  - Автоматическая стандартизация входных и целевых переменных.
  - Автоматический выбор и использование валидационной выборки.
  - Автоматическая валидация на данных, не использованных для обучения, для ранней остановки обучения в целях предотвращения переобучения.
  - Поддержка интеллектуальной автоматической настройки параметров модели.
  - Автоматическое создание скорингового кода на языке SAS для применения модели.
- Машина опорных векторов:
  - Моделирование бинарных целевых переменных.
  - Поддержка линейных и полиномиальных ядер для обучения модели.
  - Возможность включения в модель постоянных и категориальных входных признаков.
  - Автоматическое масштабирование входных признаков.
  - Возможность использования метода внутренней точки и метода активного множества.
  - Поддержка разбиения данных для проверки модели.
  - Поддержка кросс-валидации для подбора величины штрафа.
  - Автоматическое создание скорингового кода на языке SAS для применения модели.
- Факторизационные машины:
  - Поддержка разработки рекомендательных систем на основе разреженных матриц идентификаторов пользователей и рейтингов элементов.
  - Возможность применения полного тензорного разложения на попарные взаимодействия.
  - Включение дополнительных категориальных и числовых входных признаков для создания более точных моделей.
  - Возможность дополнить модели временны́ми отсечками, демографическими данными и контекстной информацией.
  - Поддержка «быстрого перезапуска» (обновление моделей на основе новых транзакций без полного переобучения).
  - Автоматическое создание скорингового кода на языке SAS для применения модели.
- Байесовские сети:
  - Выявление и обучение различных структур байесовских сетей, включая наивные модели, усиленные деревом наивные модели (TAN), усиленные байесовской сетью наивные модели (BAN), сети с отношениями «родитель — потомок» и покрытие Маркова (Markov blanket).
  - Эффективный выбор переменных с помощью тестирования на независимость.
  - Автоматический выбор оптимальной модели на основе заданных параметров.
  - Создание кода SAS или аналитического хранилища (analytic store) для скоринга данных.
  - Загрузка данных из нескольких узлов кластера и параллельное выполнение вычислений.

Аналитическая подготовка данных

- Процедуры распределенного управления данными в графическом интерфейсе.
- Масштабное исследование и описание данных.

Встроенные инструменты анализа текста

Данное решение, изначально ориентированное на анализ больших данных, позволяет изучать огромные массивы текстовых документов. Аналитик может исследовать все текстовые данные, а не только их подвыборку, что помогает получать новую полезную информацию по неизвестным темам и связям. Объединение структурированных данных с текстовыми помогает выявить ранее упущенные взаимосвязи и повышает качество аналитических прогнозных моделей.

Оценка моделей и скоринг

Для быстрого поиска наилучших моделей проверяйте различные методы моделирования в один проход и сравнивайте результаты работы нескольких обучающих алгоритмов с помощью единообразных тестов. Имплементируйте аналитическую модель как в распределенной, так и в традиционной среде с помощью автоматически созданного скорингового кода SAS.

Высокая доступность и поддержка облачных технологий

Независимо от выбранного языка программирования — Python, R, Java или Lua — специалисты по моделированию и анализу данных могут получить доступ к возможностям SAS из привычной для них среды разработки. Благодаря интерфейсам REST API для SAS Viya можно интегрировать возможности SAS с другими приложениями.

Для развертывания решения SAS Visual Data Mining and Machine Learning можно выбрать именно тот способ, который является оптимальным в соответствии с потребностями компании: локальное развертывание, развертывание в частном облаке с использованием Cloud Foundry или в публичном облаке (например, Amazon Web Services или Microsoft Azure). Доступ к данному решению также можно осуществлять при помощи предварительно развернутых и настроенных управляемых решений SAS по модели «ПО как услуга».

ПОДРОБНЕЕ »

Чтобы узнать больше о решении SAS Visual Data Mining and Machine Learning, загрузить технические документы, просмотреть скриншоты и ознакомиться с другими тематическими материалами, посетите сайт sas.com/vdmml.

Ключевые функции (продолжение)

- Профилирование для анализа кардинальности:
  - Крупномасштабное профилирование для источников входных данных.
  - Автоматические рекомендации по выбору типов и ролей переменных.
- Формирование выборок: поддержка случайной и стратифицированной выборки, выборки с избытком (oversampling) для редких событий.

Исследование данных, формирование признаков и уменьшение размерности данных.

- Дискретизация (binning) признаков.
- Высокопроизводительное заполнение пропущенных значений заданным значением, средним, псевдомедианным или случайным значением из числа непропущенных.
- Уменьшение размерности пространства признаков.
- Крупномасштабный метод главных компонент (PCA), в т.ч. метод скользящего окна и устойчивый метод главных компонент.
- Обучение без учителя с помощью кластерного анализа и кластеризации переменных разных типов.

Встроенные инструменты текстовой аналитики

- Встроенная поддержка 30 языков: английский, арабский, китайский, хорватский, чешский, датский, голландский, фарси, финский, французский, немецкий, греческий, иврит, хинди, индонезийский, итальянский, японский, корейский, норвежский, польский, португальский, русский, словацкий, словенский, испанский, шведский, тагальский, турецкий, тайский и вьетнамский.
- Определение частей речи для слов (более 15 частей речи).
- Преднастроенные правила для извлечения сущностей нескольких стандартных типов (например, местоположение, время, дата и адрес).
- Выявление в тексте групп существительных и многословных терминов. Возможность использовать их в виде единого термина при машинном обучении.
- Определение основ/стемов терминов.
- Выявление в тексте синонимичных вариантов написания терминов.
- Различные методы назначения терминам весовых коэффициентов для устранения негативных эффектов от высокой частотности терминов.
- Весовые коэффициенты можно использовать для выделения наиболее важных терминов в коллекции документов.
- Возможность использования преднастроенных старт- и стоп-листов в лингвистическом анализе и в других видах анализа.
- Редактирование списков многословных терминов, а также старт- и стоп-листов.
- Построение матричного числового представления коллекций документов с помощью тематического моделирования на основе машинного обучения.
- Семантически однородные темы, извлеченные из текстов, можно использовать как входные данные для моделей машинного обучения.
- Возможна автоматическая генерация кода для скоринга документов в SAS, включая их предобработку и лингвистический анализ.

Оценка модели

- Автоматический расчет показателей качества моделей обучения с учителем.
- Формирование статистики для интервальных и категориальных целевых переменных.
- Создание таблицы с показателями Lift для интервальной и категориальной целевых переменных.
- Создание таблицы с показателями площади под ROC-кривой для категориальной целевой модели.

Скоринг

- Автоматическое создание кода шага данных SAS для скоринга моделью.
- Применение скоринговых правил к обучающей и отложенной выборкам, а также к новым данным.

Чтобы узнать контактную информацию локального представительства SAS, посетите сайт sas.com/offices.

THE POWER TO KNOW.