CRISP-DM 1.0

Guia de mineração de dados passo a passo

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR),
Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler),
Colin Shearer (SPSS) e Rüdiger Wirth (DaimlerChrysler)

Este documento descreve o modelo de processo CRISP-DM e contém informações sobre a metodologia CRISP-DM, o modelo de referência CRISP-DM, o guia do usuário CRISP-DM e os relatórios CRISP-DM, bem como um apêndice com informações adicionais relacionadas. Este documento e as informações aqui contidas são de propriedade exclusiva dos parceiros do consórcio CRISP-DM: NCR Systems Engineering Copenhagen (EUA e Dinamarca), DaimlerChrysler AG (Alemanha), SPSS Inc. (EUA) e OHRA Verzekeringen en Bank Groep BV (Os Países Baixos).

Prefácio

O CRISP-DM foi concebido no final de 1996 por três “veteranos” do jovem e imaturo mercado de mineração de dados. A DaimlerChrysler (então Daimler-Benz) já estava à frente da maioria das organizações industriais e comerciais na aplicação de mineração de dados em suas operações comerciais. A SPSS (então ISL) forneceu serviços baseados em mineração de dados desde 1990 e lançou a primeira bancada comercial de mineração de dados — Clementine® — em 1994. A NCR, como parte de seu objetivo de agregar valor aos seus clientes de data warehouse Teradata®, estabeleceu equipes de consultores de mineração de dados e especialistas em tecnologia para atender às necessidades de seus clientes.

Naquela época, o interesse inicial do mercado em mineração de dados mostrava sinais de explosão em aceitação generalizada. Isso foi emocionante e aterrorizante. Todos nós havíamos desenvolvido nossas abordagens para mineração de dados à medida que avançávamos. Estávamos fazendo certo? Todo novo adotante de mineração de dados teria que aprender, como fizemos inicialmente, por tentativa e erro? E, do ponto de vista do fornecedor, como poderíamos demonstrar a clientes em potencial que a mineração de dados estava suficientemente madura para ser adotada como parte essencial de seus processos de negócios?

Um modelo de processo padrão, pensamos, não proprietário e disponível gratuitamente, abordaria essas questões para nós e para todos os profissionais.

Um ano depois, formamos um consórcio, inventamos um acrônimo (CRoss-Industry Standard Process for Data Mining), obtivemos financiamento da Comissão Européia e começamos a definir nossas ideias iniciais. Como o CRISP-DM foi concebido para ser neutro em termos de indústria, ferramenta e aplicativo, sabíamos que tínhamos que obter informações da maior variedade possível de profissionais e outros (como fornecedores de data warehouse e consultorias de gerenciamento) com interesse em mineração de dados. Fizemos isso criando o Grupo de Interesse Especial CRISP-DM (“The SIG”, como ficou conhecido). Lançamos o SIG transmitindo um convite às partes interessadas para se juntarem a nós em Amsterdã para um workshop de um dia inteiro: Compartilharíamos nossas ideias, os convidaríamos a apresentar as suas e discutiríamos abertamente como levar o CRISP-DM adiante.

No dia do workshop, havia um sentimento de apreensão entre os consorciados. Ninguém estaria interessado o suficiente para aparecer? Ou, se o fizessem, eles nos diriam que realmente não viam uma necessidade imperiosa de um processo padrão? Ou que nossas ideias estavam tão fora de sintonia com as dos outros que qualquer ideia de padronização era uma fantasia impraticável?
A oficina superou todas as nossas expectativas. Três coisas se destacaram: ÿ Duas
vezes mais pessoas do que esperávamos inicialmente ÿ Houve um consenso
esmagador de que a indústria precisava de um processo padrão e precisava dele agora ÿ À medida que os participantes
apresentavam suas opiniões sobre mineração de dados a partir de sua experiência em projetos, tornou-se claro que, embora houvesse diferenças
superficiais - principalmente na demarcação de fases e na terminologia - havia um enorme terreno comum em como eles viam o processo de mineração de dados

Ao final do workshop, nos sentimos confiantes de que poderíamos fornecer, com a contribuição e a crítica do SIG, um modelo de processo padrão para atender à comunidade de mineração de dados.

Nos dois anos e meio seguintes, trabalhamos para desenvolver e refinar o CRISP-DM. Fizemos testes em projetos de mineração de dados em grande escala ao vivo na Mercedes-Benz e em nosso parceiro do setor de seguros, OHRA. Trabalhamos na integração do CRISP-DM com ferramentas comerciais de mineração de dados. O SIG provou ser inestimável, crescendo para mais de 200 membros e realizando workshops em Londres, Nova York e Bruxelas.

Ao final da parte do projeto financiada pela CE - meados de 1999 - produzimos o que consideramos um rascunho de boa qualidade do modelo de processo. Aqueles familiarizados com esse rascunho descobrirão que um ano depois, embora agora muito mais completo e melhor apresentado, o CRISP-DM 1.0 não é radicalmente diferente. Estávamos cientes de que, durante o projeto, o modelo de processo ainda era um trabalho em andamento; O CRISP-DM só foi validado em um conjunto restrito de projetos. No ano passado, a DaimlerChrysler teve a oportunidade de aplicar o CRISP-DM a uma gama mais ampla de aplicações. Os grupos de serviços profissionais da SPSS e da NCR adotaram o CRISP-DM e o usaram com sucesso em vários contratos com clientes que abrangem muitos setores e problemas de negócios.
Ao longo desse tempo, vimos fornecedores de serviços de fora do consórcio adotarem o CRISP-DM, repetidas referências a ele por analistas como o padrão de fato para o setor e uma crescente conscientização de sua importância entre os clientes (CRISP-DM agora é frequentemente referenciado em editais de licitação e em documentos de RFP). Acreditamos que nossa iniciativa foi totalmente justificada e, embora futuras extensões e melhorias sejam desejáveis e inevitáveis, consideramos o CRISP-DM Versão 1.0 suficientemente validado para ser publicado e distribuído.

O CRISP-DM não foi construído de maneira teórica e acadêmica, trabalhando a partir de princípios técnicos, nem comitês de elite de gurus o criaram a portas fechadas. Ambas as abordagens para o desenvolvimento de metodologias foram tentadas no passado, mas raramente levaram a padrões práticos, bem-sucedidos e amplamente adotados. O CRISP-DM é bem-sucedido porque é solidamente baseado na experiência prática e real de como as pessoas conduzem projetos de mineração de dados. E, a esse respeito, somos imensamente gratos aos muitos profissionais que contribuíram com seus esforços e ideias ao longo do projeto.

O consórcio CRISP-DM

agosto de 2000

Índice

Introdução

A metodologia CRISP-DM

1.1 Divisão hierárquica

A metodologia CRISP-DM é descrita em termos de um modelo de processo hierárquico, composto por conjuntos de tarefas descritas em quatro níveis de abstração (do geral ao específico): fase, tarefa genérica, tarefa especializada e instância do processo (ver figura 1).

No nível superior, o processo de mineração de dados é organizado em várias fases; cada fase consiste em várias tarefas genéricas de segundo nível. Este segundo nível é chamado genérico porque pretende ser geral o suficiente para cobrir todas as situações possíveis de mineração de dados. As tarefas genéricas devem ser tão completas e estáveis quanto possível. Completo significa abranger todo o processo de mineração de dados e todas as possíveis aplicações de mineração de dados. Estável significa que o modelo deve ser válido para desenvolvimentos imprevistos, como novas técnicas de modelagem.

O terceiro nível, o nível de tarefa especializada, é o lugar para descrever como as ações nas tarefas genéricas devem ser executadas em determinadas situações específicas. Por exemplo, no segundo nível pode haver uma tarefa genérica chamada limpar dados. O terceiro nível descreve como essa tarefa difere em diferentes situações, como limpeza de valores numéricos versus limpeza de valores categóricos ou se o tipo de problema é agrupamento ou modelagem preditiva.

A descrição das fases e tarefas como etapas discretas executadas em uma ordem específica representa uma sequência idealizada de eventos. Na prática, muitas das tarefas podem ser executadas em uma ordem diferente e muitas vezes será necessário voltar repetidamente às tarefas anteriores e repetir certas ações. Nosso modelo de processo não tenta capturar todas essas rotas possíveis por meio do processo de mineração de dados porque isso exigiria um modelo de processo excessivamente complexo.

O quarto nível, a instância do processo, é um registro das ações, decisões e resultados de uma mineração de dados real. noivado. Uma instância de processo é organizado de acordo com as tarefas definidas nos níveis superiores, mas representa o que realmente aconteceu em um determinado engajamento, ao invés do que acontece em geral.

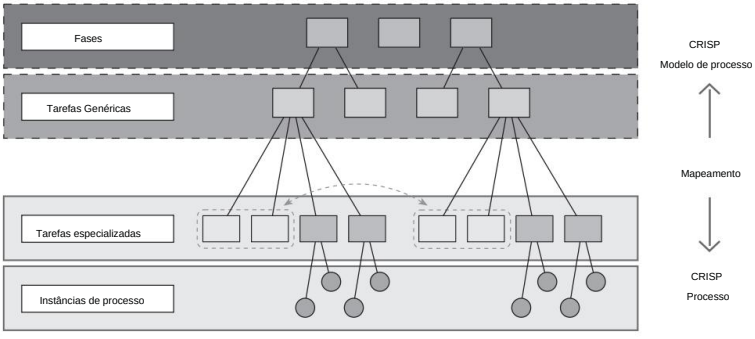

Figura 1: Divisão em quatro níveis da metodologia CRISP-DM

1.2 Modelo de referência e guia do usuário

Horizontalmente, a metodologia CRISP-DM distingue entre o modelo de referência e o guia do usuário. O modelo de referência apresenta uma visão geral rápida das fases, tarefas e suas saídas e descreve o que fazer em um projeto de mineração de dados. O guia do usuário fornece dicas e sugestões mais detalhadas para cada fase e cada tarefa dentro de uma fase e descreve como realizar uma mineração de dados projeto.

Este documento abrange tanto o modelo de referência quanto o guia do usuário em nível genérico.

Mapeamento de modelos genéricos para modelos especializados

2.1 Contexto de mineração de dados

O contexto de mineração de dados direciona o mapeamento entre o nível genérico e especializado no CRISP-DM. Atualmente, distinguimos quatro dimensões diferentes de contextos de mineração de dados: ÿ O **domínio do aplicativo** é a área específica na qual o projeto de mineração de dados ocorre ÿ O **tipo de problema de mineração de dados** descreve a(s) classe(s) específica(s) de objetivo(s) que os dados trata do projeto de mineração (ver também Apêndice 2)

ÿ O **aspecto técnico** abrange questões específicas na mineração de dados que descrevem diferentes desafios (técnicos) que geralmente ocorrem durante a mineração de dados

ÿ A dimensão de **ferramenta e técnica** especifica quais ferramentas e/ou técnicas de mineração de dados são aplicadas durante os dados projeto de mineração

A Tabela 1 abaixo resume essas dimensões dos contextos de mineração de dados e mostra exemplos específicos para cada dimensão.

Tabela 1: Dimensões dos contextos e exemplos de mineração de dados

	Contexto de mineração de dados			
Dimensão	**Aplicativo Domínio**	**Mineração de dados Tipo de problema**	**Técnico Aspecto**	**Ferramenta e Técnica**
Exemplos	Resposta Modelagem	Descrição e resumo	Ausente valores	clementina
	Churn Predição	Segmentação	Outliers	MineSet
	...	Conceito Descrição	...	Decisão Árvore
		Classificação		...
		Predição		
		Dependência Análise		

Um contexto específico de mineração de dados é um valor concreto para uma ou mais dessas dimensões. Por exemplo, um projeto de mineração de dados lidando com um problema de classificação na previsão de churn constitui um contexto específico. Quanto mais valores para diferentes dimensões de contexto forem fixados, mais concreto será o contexto de mineração de dados.

2.2 Mapeamentos com contextos

Distinguimos entre dois tipos diferentes de mapeamento entre nível genérico e especializado no CRISP-DM.

Mapeamento para o presente: se aplicarmos apenas o modelo de processo genérico para executar um único projeto de mineração de dados e tentarmos mapear tarefas genéricas e suas descrições para o projeto específico conforme necessário, falamos de um único mapeamento para (provavelmente) apenas um uso .

Mapeamento para o futuro: se especializarmos sistematicamente o modelo de processo genérico de acordo com um contexto predefinido (ou analisarmos e consolidarmos sistematicamente experiências de um único projeto em direção a um modelo de processo especializado para uso futuro em contextos comparáveis), falamos sobre escrever explicitamente um modelo de processo especializado em termos de CRISP-DM.

Qual tipo de mapeamento é apropriado para seus próprios propósitos depende de seu contexto de mineração de dados específico e das necessidades de sua organização.

2.3 Como mapear

A estratégia básica para mapear o modelo de processo genérico para o nível especializado é a mesma para ambos os tipos de mapeamentos: ÿ
Analisar seu contexto específico ÿ
Remover todos os detalhes não aplicáveis ao seu contexto ÿ
Adicionar detalhes específicos ao seu contexto ÿ
Especializar (ou instanciar) conteúdos genéricos de acordo com características concretas de seu contexto ÿ
Possivelmente renomeie conteúdos genéricos para fornecer significados mais explícitos em seu contexto por uma questão de clareza

Descrição das peças

3.1 Conteúdo

O modelo de processo CRISP-DM (este documento) está organizado em cinco partes diferentes: ÿ A Parte
I é esta introdução à metodologia CRISP-DM, que fornece algumas diretrizes gerais para mapear os processos genéricos
modelo de processo para modelos de processo especializados
ÿ A Parte II descreve o modelo de referência CRISP-DM, suas fases, tarefas genéricas e saídas ÿ A Parte III
apresenta o guia do usuário CRISP-DM, que vai além da descrição pura de fases, tarefas genéricas e saídas,
e contém conselhos mais detalhados sobre como realizar projetos de mineração de dados
ÿ A Parte IV enfoca os relatórios a serem produzidos durante e após um projeto e sugere esboços para esses relatórios. Isso também
mostra referências cruzadas entre saídas e tarefas.
ÿ A Parte V é o apêndice, que inclui um glossário de terminologia importante e uma caracterização da mineração de dados
tipos de problema

3.2 Finalidade

Os usuários e leitores deste documento devem estar cientes das seguintes instruções: ÿ Se você está
lendo o modelo de processo CRISP-DM pela primeira vez, comece com a parte I, a introdução, para entender a metodologia CRISP-DM, todos seus conceitos
e como os diferentes conceitos se relacionam entre si. Em leituras posteriores, você pode pular a introdução e retornar a ela apenas se necessário para
esclarecimentos.
ÿ Se você precisar de acesso rápido a uma visão geral do modelo de processo CRISP-DM, consulte a parte II, o modelo de referência CRISP-DM, ou
para iniciar um projeto de mineração de dados rapidamente ou para obter uma introdução ao guia do usuário CRISP-
DM ÿ Se você precisar de orientação detalhada na execução de seu projeto de mineração de dados, a parte III, o guia do usuário CRISP-DM, é o mais valioso
parte deste documento. Observação: se você não leu a introdução ou o modelo de referência primeiro, volte e leia-os primeiro
duas partes.
ÿ Se você estiver no estágio de mineração de dados ao redigir seus relatórios, vá para a parte IV. Se você preferir gerar sua entrega
descrições durante o projeto, vá e volte entre as partes III e IV conforme desejado.
ÿ Por fim, o apêndice é útil como informações básicas adicionais sobre CRISP-DM e mineração de dados. Use o apêndice para
procure vários termos se você ainda não for um especialista na área.

II O modelo de referência CRISP-DM

O modelo de processo atual para mineração de dados fornece uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases de um projeto, suas respectivas tarefas e os relacionamentos entre essas tarefas. Nesse nível de descrição, não é possível identificar todos os relacionamentos. Podem existir relacionamentos entre quaisquer tarefas de mineração de dados, dependendo dos objetivos, do histórico e do interesse do usuário – e o mais importante – nos dados.

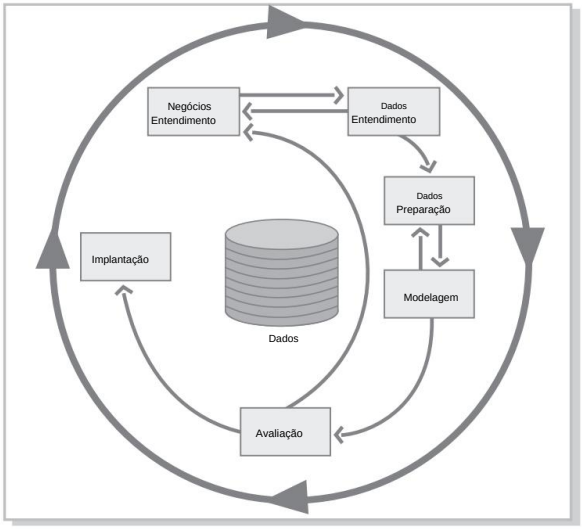

Figura 2: Fases do modelo de referência CRISP-DM

O ciclo de vida de um projeto de mineração de dados consiste em seis fases, mostradas na **Figura** 2. A sequência das fases não é rígida. Mover-se para frente e para trás entre diferentes fases é sempre necessário. O resultado de cada fase determina qual fase, ou tarefa específica de uma fase, deve ser executada a seguir. As setas indicam as dependências mais importantes e frequentes entre as fases.

O círculo externo na Figura 2 simboliza a natureza cíclica da própria mineração de dados. A mineração de dados não termina quando uma solução é implantada. As lições aprendidas durante o processo e com a solução implantada podem desencadear novas questões de negócios, muitas vezes mais focadas. Os processos subsequentes de mineração de dados se beneficiarão das experiências dos anteriores. A seguir, descrevemos brevemente cada fase:

Compreensão do negócio

Esta fase inicial concentra-se na compreensão dos objetivos e requisitos do projeto de uma perspectiva de negócios, convertendo esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para atingir os objetivos.

Compreensão dos

dados A fase de compreensão dos dados começa com a coleta inicial de dados e prossegue com atividades que permitem que você se familiarize com os dados, identifique problemas de qualidade de dados, descubra os primeiros insights sobre os dados e/ou detecte subconjuntos interessantes para formar hipóteses sobre informações ocultas.

Preparação de dados

A fase de preparação de dados abrange todas as atividades necessárias para construir o conjunto de dados final [dados que serão alimentados na(s) ferramenta(s) de modelagem] a partir dos dados brutos iniciais. As tarefas de preparação de dados provavelmente serão executadas várias vezes e não em qualquer ordem prescrita. As tarefas incluem seleção de tabela, registro e atributo, bem como transformação e limpeza de dados para ferramentas de modelagem.

Modelagem

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos específicos na forma de dados. Portanto, muitas vezes é necessário voltar à fase de preparação de dados.

Avaliação

Nesta fase do projeto, você construiu um modelo (ou modelos) que parece ter alta qualidade do ponto de vista da análise de dados. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo minuciosamente e revisar as etapas executadas para criá-lo, para ter certeza de que o modelo atende adequadamente aos objetivos do negócio. Um dos principais objetivos é determinar se há alguma questão comercial importante que não foi suficientemente considerada. No final desta fase, deve ser tomada uma decisão sobre a utilização dos resultados da mineração de dados.

Implantação A

criação do modelo geralmente não é o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente possa utilizá-lo. Frequentemente, envolve a aplicação de modelos “ao vivo” nos processos de tomada de decisão de uma organização – por exemplo, personalização em tempo real de páginas da Web ou pontuação repetida de bancos de dados de marketing. Dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível em toda a empresa. Em muitos casos, é o cliente, não o analista de dados, quem realiza as etapas de implantação. No entanto, mesmo que o trabalho de implantação seja feito pelo analista, é importante que o cliente entenda de antemão quais ações precisam ser realizadas para de fato fazer uso dos modelos criados.

A Figura 3 apresenta um esboço das fases acompanhadas de tarefas genéricas (negrito) e saídas (itálico). Nas seções a seguir, descrevemos cada tarefa genérica e suas saídas com mais detalhes. Concentramos nossa atenção em visões gerais de tarefas e resumos de resultados.

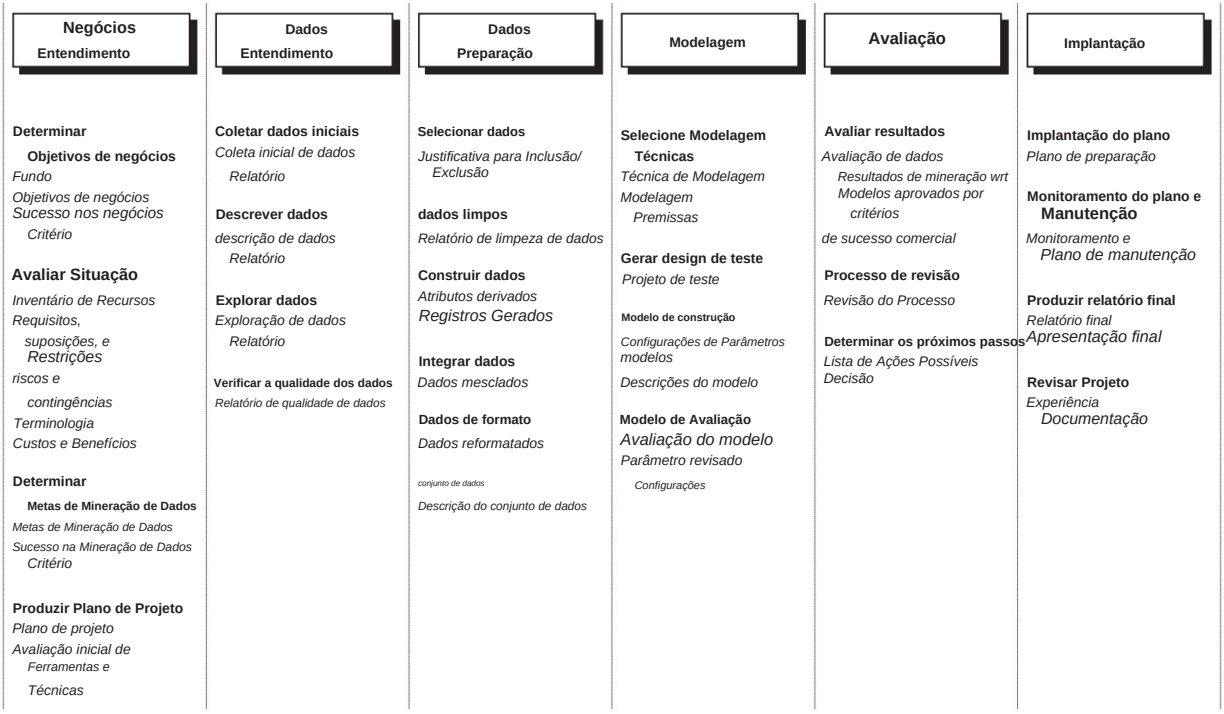

Figura 3: Tarefas genéricas (negrito) e saídas (itálico) do modelo de referência CRISP-DM

1 Compreensão do negócio

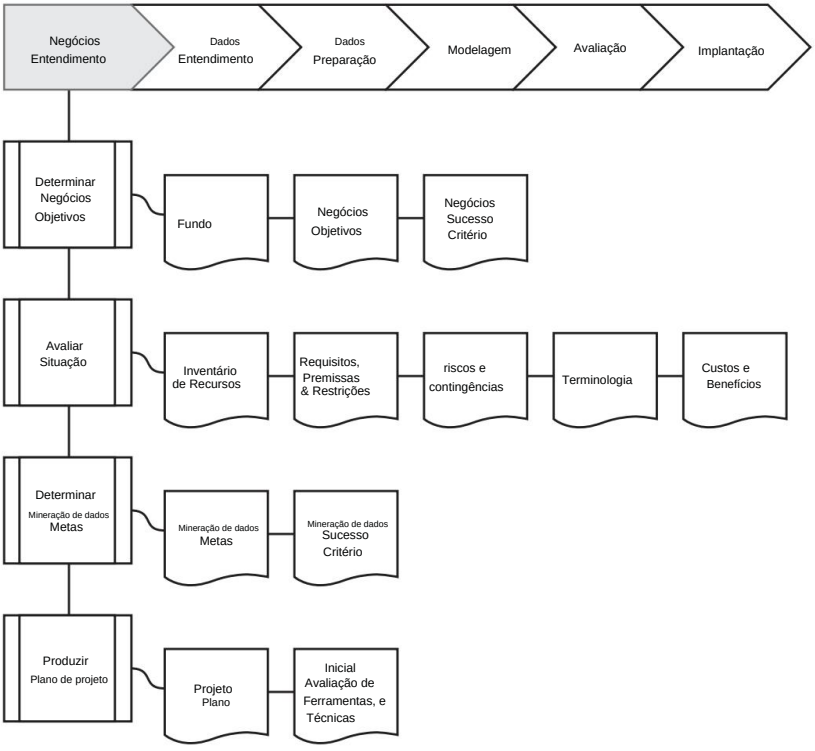

Figura 4: Entendimento do Negócio

1.1 Determinar os objetivos de negócios

Tarefa

Determinar os objetivos de negócios

O primeiro objetivo do analista de dados é entender completamente, de uma perspectiva de negócios, o que o cliente realmente deseja realizar. Freqüentemente, o cliente tem muitos objetivos e restrições conflitantes que devem ser devidamente equilibrados. O objetivo do analista é descobrir fatores importantes, no início, que podem influenciar o resultado do projeto. Uma possível consequência de negligenciar esta etapa é gastar muito esforço produzindo as respostas certas para as perguntas erradas.

Saídas

Histórico

Registre as informações conhecidas sobre a situação comercial da organização no início do projeto.

Objetivos de negócios

Descreva o objetivo principal do cliente, de uma perspectiva de negócios. Além do objetivo comercial principal, normalmente há outras questões comerciais relacionadas que o cliente gostaria de abordar. Por exemplo, o principal objetivo de negócios pode ser manter os clientes atuais, prevendo quando eles estão propensos a mudar para um concorrente. Exemplos de questões de negócios relacionadas são “Como o canal principal usado (por exemplo, caixa eletrônico, visita à agência, Internet) afeta a permanência ou saída dos clientes?” ou “Reduzirá significativamente as taxas do caixa eletrônico
reduzir o número de clientes de alto valor que saem?”

Critérios de sucesso do negócio

Descreva os critérios para um resultado bem-sucedido ou útil para o projeto do ponto de vista do negócio. Isso pode ser bastante específico e passível de ser medido objetivamente, por exemplo, redução da rotatividade de clientes a um determinado nível, ou pode ser geral e subjetivo, como “dar informações úteis sobre os relacionamentos”. Neste último caso, deve ser indicado quem faz o julgamento subjetivo.

1.2 Avalie a situação

Tarefa

Avalie a situação

Essa tarefa envolve uma descoberta de fatos mais detalhada sobre todos os recursos, restrições, suposições e outros fatores que devem ser considerados na determinação do objetivo da análise de dados e do plano do projeto. Na tarefa anterior, seu objetivo é chegar rapidamente ao cerne da situação. Aqui, você deseja expandir
os detalhes.

Saídas

Inventário de recursos

Liste os recursos disponíveis para o projeto, incluindo pessoal (especialistas em negócios, especialistas em dados, suporte técnico, especialistas em mineração de dados), dados (extratos fixos, acesso a dados ativos, armazenados ou operacionais), recursos de computação (plataformas de hardware) e software (ferramentas de mineração de dados, outro software relevante).

Requisitos, premissas e restrições Liste todos os requisitos do projeto, incluindo cronograma de conclusão, compreensibilidade e qualidade dos resultados e segurança, bem como questões legais. Como parte desta saída, certifique-se de que você tem permissão para usar os dados.

Liste as suposições feitas pelo projeto. Podem ser suposições sobre os dados que podem ser verificados durante a mineração de dados, mas também podem incluir suposições não verificáveis sobre os negócios relacionados ao projeto. É particularmente importante listar o último caso isso afete a validade dos resultados.

Liste as restrições do projeto. Podem ser restrições na disponibilidade de recursos, mas também podem incluir restrições tecnológicas, como o tamanho do conjunto de dados que é prático usar para modelagem.

Riscos e contingências

Liste os riscos ou eventos que podem atrasar o projeto ou causar seu fracasso. Liste os planos de contingência correspondentes, que ação será tomada se esses riscos ou eventos ocorrerem.

Terminologia

Compile um glossário de terminologia relevante para o projeto. Isso pode incluir dois componentes:

(1) Um glossário de terminologia de negócios relevante, que faz parte do entendimento de negócios disponível para o projeto. Construir este glossário é uma útil “elicitação de conhecimento” e exercício de educação.

(2) Um glossário de terminologia de mineração de dados, ilustrado com exemplos relevantes para o negócio problema em questão

Custos e benefícios

Construir uma análise de custo-benefício para o projeto, que compara os custos do projeto com os benefícios potenciais para o negócio se for bem-sucedido. A comparação deve ser o mais específica possível. Por exemplo, use medidas monetárias em uma situação comercial.

1.3 Determinar metas de mineração de dados

Tarefa

Determinar metas de mineração de dados Uma meta de negócios declara os objetivos na terminologia de negócios. Uma meta de mineração de dados declara os objetivos do projeto em termos técnicos. Por exemplo, a meta de negócios pode ser “Aumentar as vendas do catálogo para clientes existentes”. Uma meta de mineração de dados pode ser “Prever quantos widgets um cliente comprará, considerando suas compras nos últimos três anos, informações demográficas (idade, salário, cidade etc.) e o preço do item”.

Saídas

Metas de mineração de dados Descreva as saídas pretendidas do projeto que permitem a realização dos objetivos de negócios.

Critérios de sucesso da mineração de dados Defina os critérios para um resultado bem-sucedido para o projeto em termos técnicos, por exemplo, um certo nível de precisão preditiva ou um perfil de propensão a comprar com um determinado grau de “elevação”. Tal como acontece com os critérios de sucesso empresarial, pode ser necessário descrevê-los em termos subjetivos, caso em que a pessoa ou pessoas que fazem o julgamento subjetivo devem ser identificadas.

1.4 Produzir plano de projeto

Tarefa

Produzir plano de projeto

Descreva o plano pretendido para atingir as metas de mineração de dados e, assim, atingir as metas de negócios.
O plano deve especificar as etapas a serem executadas durante o restante do projeto, incluindo a seleção inicial de ferramentas e técnicas.

Saídas

Plano do projeto Liste as etapas a serem executadas no projeto, juntamente com sua duração, recursos necessários, entradas, saídas e dependências. Sempre que possível, torne explícitas as iterações em larga escala no processo de mineração de dados, por exemplo, repetições das fases de modelagem e avaliação.

Como parte do plano do projeto, também é importante analisar as dependências entre o cronograma e os riscos.
Marque os resultados dessas análises explicitamente no plano do projeto, de preferência com ações e recomendações se os riscos se manifestam.

Nota: o plano do projeto contém planos detalhados para cada fase. Decida neste ponto qual estratégia de avaliação será usada na fase de avaliação.

O plano do projeto é um documento dinâmico no sentido de que, ao final de cada fase, é necessária uma revisão do progresso e das realizações e uma atualização correspondente do plano do projeto é recomendada. Pontos de revisão específicos para essas atualizações fazem parte do plano do projeto.

Avaliação inicial de ferramentas e técnicas

Ao final da primeira fase, uma avaliação inicial de ferramentas e técnicas deve ser realizada. Aqui, por exemplo, você seleciona uma ferramenta de mineração de dados que oferece suporte a vários métodos para diferentes estágios do processo. É importante avaliar ferramentas e técnicas no início do processo, pois a seleção de ferramentas e técnicas pode influenciar todo o projeto.

2 Entendimento de dados

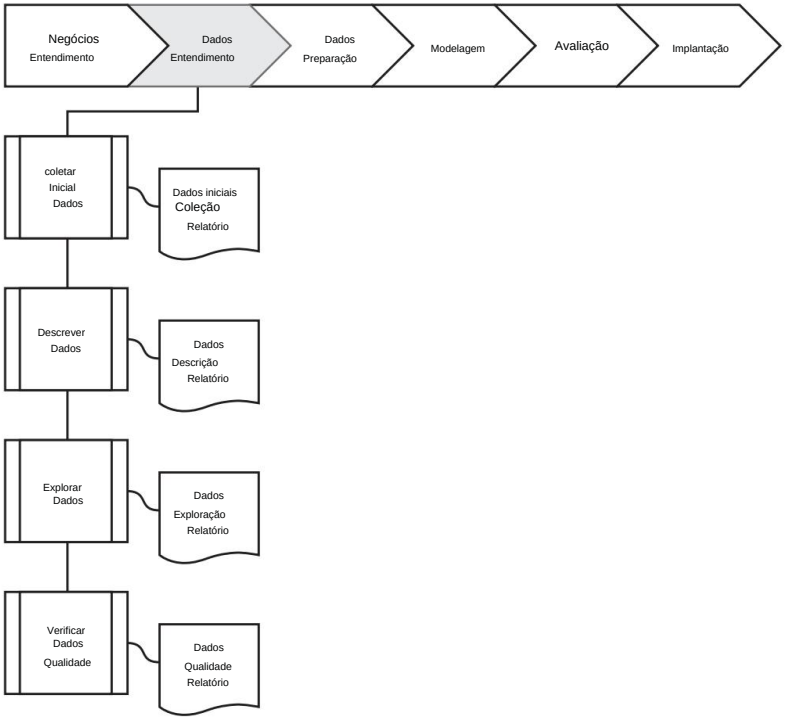

Figura 5: Compreensão dos dados

2.1 Coletar dados iniciais

Tarefa **Coletar dados iniciais**

Adquira os dados (ou acesso aos dados) listados nos recursos do projeto. Essa coleta inicial inclui o carregamento de dados, se necessário para a compreensão dos dados. Por exemplo, se você utiliza uma ferramenta específica para entendimento de dados, faz todo o sentido carregar seus dados nesta ferramenta. Esse esforço possivelmente leva a etapas iniciais de preparação de dados.

Observação: se você adquirir várias fontes de dados, a integração é um problema adicional, aqui ou na fase posterior de preparação de dados.

Saída **Relatório inicial de coleta de dados**

Liste os conjuntos de dados adquiridos, juntamente com suas localizações, os métodos usados para adquiri-los e quaisquer problemas encontrados. Registre os problemas encontrados e quaisquer resoluções alcançadas. Isso ajudará na replicação futura deste projeto ou na execução de projetos futuros semelhantes.

2.2 Descrever dados

Tarefa **Descrever dados**

Examine as propriedades “brutas” ou “superficiais” dos dados adquiridos e relate os resultados.

Saída **Relatório de descrição de**

dados Descreva os dados que foram adquiridos, incluindo o formato dos dados, a quantidade de dados (por exemplo, o número de registros e campos em cada tabela), as identidades dos campos e quaisquer outras características de superfície que tenham sido descoberto. Avalie se os dados adquiridos satisfazem os requisitos relevantes.

2.3 Explorar dados

Tarefa **Explorar dados**

Esta tarefa aborda questões de mineração de dados usando técnicas de consulta, visualização e geração de relatórios. Isso inclui distribuição de atributos-chave (por exemplo, o atributo de destino de uma tarefa de previsão), relacionamentos entre pares ou pequenos números de atributos, resultados de agregações simples, propriedades de subpopulações significativas e análises estatísticas simples. Essas análises podem abordar diretamente os objetivos de mineração de dados; eles também podem contribuir ou refinar a descrição de dados e relatórios de qualidade, e alimentar a transformação e outras etapas de preparação de dados necessárias para análise posterior.

Saída	**Relatório de exploração de dados** Descreva os resultados desta tarefa, incluindo as primeiras descobertas ou hipóteses iniciais e seu impacto no restante do projeto. Se apropriado, inclua gráficos e plotagens para indicar características de dados que sugiram um exame mais aprofundado de subconjuntos de dados interessantes.

2.4 Verifique a qualidade dos dados

Tarefa	**Verifique a qualidade dos dados** Examine a qualidade dos dados, abordando questões como: Os dados estão completos (abrangem todos os casos necessários)? Está correto ou contém erros e, se houver erros, quão comuns eles são? Há valores ausentes nos dados? Em caso afirmativo, como são representados, onde ocorrem e quão comuns são?
Saída	**Relatório de qualidade dos dados** Liste os resultados da verificação da qualidade dos dados; se houver problemas de qualidade, liste as soluções possíveis. As soluções para problemas de qualidade de dados geralmente dependem muito dos dados e do conhecimento do negócio.

3 Preparação de dados

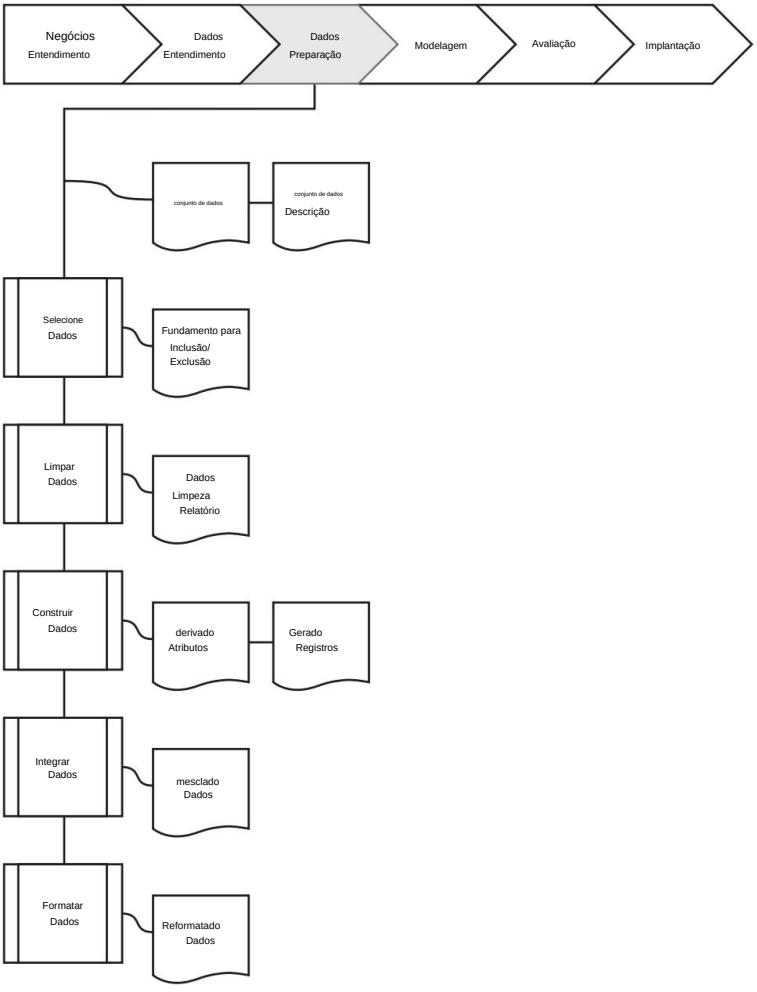

Figura 6: preparação de dados

Saídas

conjunto de dados

Trata-se do(s) conjunto(s) de dados produzido(s) pela fase de preparação de dados, que serão utilizados para modelagem ou para o trabalho de análise principal do projeto.

Descrição do conjunto de dados Descreva o(s) conjunto(s) de dados que serão usados para a modelagem e o trabalho de análise principal do projeto.

3.1 Selecione os dados

Tarefa

Selecionar dados

Decidir sobre os dados a serem usados para análise. Os critérios incluem relevância para as metas de mineração de dados, qualidade e restrições técnicas, como limites de volume ou tipos de dados. Observe que a seleção de dados abrange a seleção de atributos (colunas), bem como a seleção de registros (linhas) em uma tabela.

Saída

Justificativa da inclusão/exclusão Liste os dados a serem incluídos/excluídos e os motivos dessas decisões.

3.2 Limpar dados

Tarefa

Limpar dados

Aumente a qualidade dos dados para o nível exigido pelas técnicas de análise selecionadas. Isso pode envolver a seleção de subconjuntos limpos dos dados, a inserção de padrões adequados ou técnicas mais ambiciosas, como a estimativa de dados ausentes por modelagem.

Saída

Relatório de limpeza de dados Descreva quais decisões e ações foram tomadas para resolver os problemas de qualidade de dados relatados durante a tarefa Verificar qualidade de dados da fase Entendimento de dados. As transformações dos dados para fins de limpeza e o possível impacto nos resultados da análise devem ser considerados.

3.3 Construir dados

Tarefa

Construir dados

Esta tarefa inclui operações construtivas de preparação de dados, como a produção de atributos derivados ou novos registros inteiros, ou valores transformados para atributos existentes.

Saídas

Atributos derivados

Atributos derivados são novos atributos construídos a partir de um ou mais atributos existentes no mesmo registro. Exemplo: área = comprimento * largura.

registros gerados

Descrever a criação de registros completamente novos. Exemplo: Crie registros para clientes que não fizeram nenhuma compra no ano passado. Não havia razão para ter tais registros nos dados brutos, mas para fins de modelagem pode fazer sentido representar explicitamente o fato de que certos clientes não fizeram nenhuma compra.

3.4 Integrar dados

Tarefa

Integrar dados

Estes são métodos pelos quais as informações são combinadas de várias tabelas ou registros para criar novos registros ou valores.

Saída

Dados

mesclados Mesclar tabelas refere-se à junção de duas ou mais tabelas que possuem informações diferentes sobre os mesmos objetos. Exemplo: uma rede varejista possui uma tabela com informações sobre as características gerais de cada loja (ex: área útil, tipo de shopping), outra tabela com dados resumidos de vendas (ex: lucro, variação percentual nas vendas do ano anterior) e outra com informações sobre a demografia da área circundante.
Cada uma dessas tabelas contém um registro para cada loja. Essas tabelas podem ser mescladas em uma nova tabela com um registro para cada loja, combinando os campos das tabelas de origem.

Dados mesclados também cobrem agregações. Agregação refere-se a operações nas quais novos valores são calculados resumindo informações de vários registros e/ou tabelas. Por exemplo, converter uma tabela de compras de clientes onde há um registro para cada compra em uma nova tabela onde há um registro para cada cliente, com campos como número de compras, valor médio da compra, percentual de pedidos debitados no cartão de crédito, porcentagem de itens em promoção, etc.

3.5 Formatar dados

Tarefa

Dados de formato

As transformações de formatação referem-se principalmente a modificações *sintáticas* feitas nos dados que não alteram seu significado, mas podem ser exigidas pela ferramenta de modelagem.

Saída

dados reformatados

Algumas ferramentas têm requisitos na ordem dos atributos, como o primeiro campo sendo um identificador exclusivo para cada registro ou o último campo sendo o campo de resultado que o modelo deve prever.

Pode ser importante alterar a ordem dos registros no conjunto de dados. Talvez a ferramenta de modelagem exija que os registros sejam classificados de acordo com o valor do atributo de resultado. Normalmente, os registros do conjunto de dados são inicialmente ordenados de alguma forma, mas o algoritmo de modelagem precisa que eles estejam em uma ordem bastante aleatória. Por exemplo, ao usar redes neurais, geralmente é melhor que os registros sejam apresentados em uma ordem aleatória, embora algumas ferramentas lidem com isso automaticamente sem intervenção explícita do usuário.

Além disso, há alterações puramente sintáticas feitas para satisfazer os requisitos da ferramenta de modelagem específica. Exemplos: remover vírgulas de campos de texto em arquivos de dados delimitados por vírgulas, aparar todos os valores máximo de 32 caracteres.

4 Modelagem

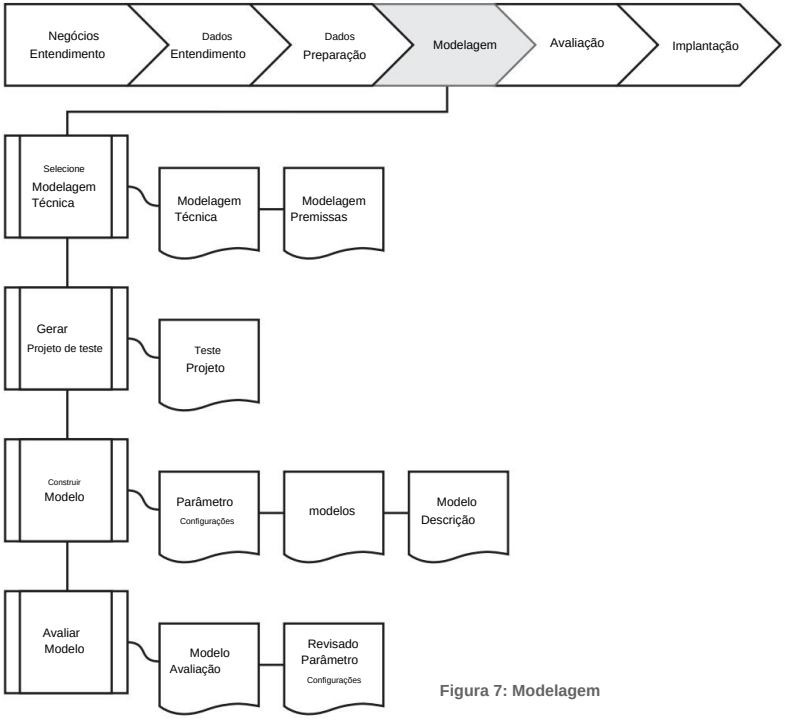

Figura 7: Modelagem

4.1 Selecione a técnica de modelagem

Tarefa **Selecione a técnica de modelagem**

Como primeira etapa na modelagem, selecione a técnica de modelagem real que será usada. Embora você já tenha selecionado uma ferramenta durante a fase de Compreensão do Negócio, esta tarefa refere-se à técnica de modelagem específica, por exemplo, construção de árvore de decisão com 5.0 ou geração de rede neural com propagação reversa. Se várias técnicas forem aplicadas, execute esta tarefa separadamente para cada técnica.

Saídas **Técnica de modelagem**

Documente a técnica de modelagem real que será usada.

Suposições de modelagem

Muitas técnicas de modelagem fazem suposições específicas sobre os dados - por exemplo, que todos os atributos têm distribuições uniformes, nenhum valor ausente permitido, atributo de classe deve ser simbólico, etc. Registre todas as suposições feitas.

4.2 Gerar projeto de teste **Gerar**

Tarefa **projeto de teste** Antes de realmente construirmos um modelo, precisamos gerar um procedimento ou mecanismo para testar a qualidade e validade do modelo. Por exemplo, em tarefas de mineração de dados supervisionadas, como classificação, é comum usar taxas de erro como medidas de qualidade para modelos de mineração de dados. Portanto, normalmente separamos o conjunto de dados em conjuntos de treinamento e teste, construímos o modelo no conjunto de treinamento e estimamos sua qualidade no conjunto de teste separado.

Saída **Projeto de teste** Descreva o plano pretendido para treinar, testar e avaliar os modelos. Um componente principal do plano é determinar como dividir o conjunto de dados disponível em conjuntos de dados de treinamento, teste e validação.

4.3 Construir modelo

Tarefa **Construir modelo**

Execute a ferramenta de modelagem no conjunto de dados preparado para criar um ou mais modelos.

Saídas **Configurações de parâmetros** Com qualquer ferramenta de modelagem, geralmente há um grande número de parâmetros que podem ser ajustados. Liste os parâmetros e seus valores escolhidos, juntamente com a justificativa para a escolha das configurações dos parâmetros.

modelos

Esses são os modelos reais produzidos pela ferramenta de modelagem, não um relatório.

Descrições do modelo

Descreva os modelos resultantes. Relate a interpretação dos modelos e documente quaisquer dificuldades encontradas com seus significados.

4.4 Avaliar o modelo

Tarefa **Avaliar modelo**

O engenheiro de mineração de dados interpreta os modelos de acordo com seu conhecimento de domínio, os critérios de sucesso da mineração de dados e o projeto de teste desejado. O engenheiro de mineração de dados julga tecnicamente o sucesso da aplicação de técnicas de modelagem e descoberta; ele entra em contato com analistas de negócios e especialistas de domínio posteriormente para discutir os resultados da mineração de dados no contexto de negócios. Observe que esta tarefa considera apenas modelos, enquanto a fase de avaliação também leva em consideração todos os outros resultados produzidos no decorrer do projeto.

O engenheiro de mineração de dados tenta classificar os modelos. Ele avalia os modelos de acordo com os critérios de avaliação. Tanto quanto possível, ele também leva em consideração os objetivos de negócios e os critérios de sucesso do negócio.
Na maioria dos projetos de mineração de dados, o engenheiro de mineração de dados aplica uma única técnica mais de uma vez ou gera resultados de mineração de dados com várias técnicas diferentes. Nessa tarefa, ele também compara todos os resultados de acordo com os critérios de avaliação.

Saídas **Avaliação do modelo**

Resuma os resultados desta tarefa, liste as qualidades dos modelos gerados (por exemplo, em termos de precisão) e classifique sua qualidade em relação uns aos outros.

Configurações de parâmetro

revisadas De acordo com a avaliação do modelo, revise as configurações de parâmetro e ajuste-as para a próxima execução na tarefa Construir modelo. Repita a construção e avaliação do modelo até que você acredite fortemente que encontrou o (s) *melhor* (es) modelo(s). Documente todas essas revisões e avaliações.

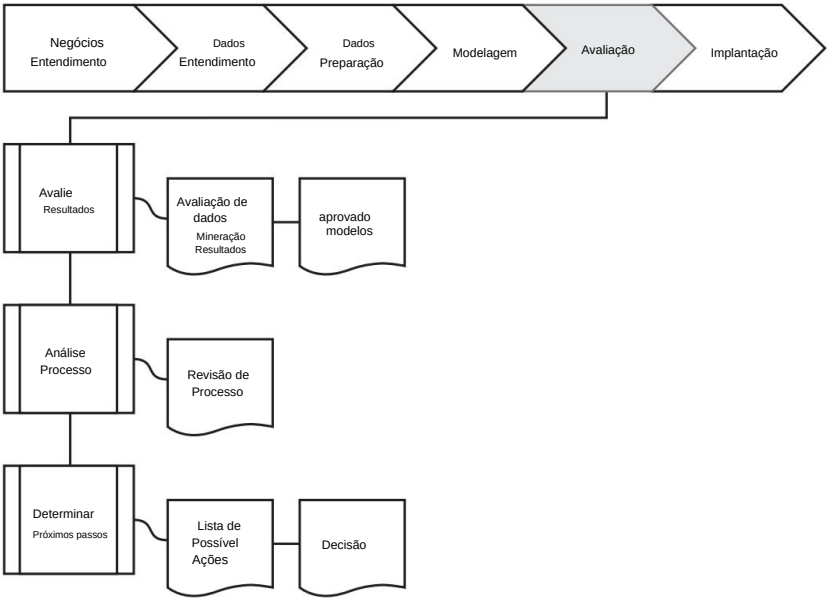

Figura 8: Avaliação

5.1 Avaliar resultados

Tarefa **Avalie os resultados**

As etapas de avaliação anteriores lidaram com fatores como a precisão e a generalidade do modelo. Esta etapa avalia o grau em que o modelo atende aos objetivos de negócios e procura determinar se há algum motivo comercial para que esse modelo seja deficiente. Outra opção é testar o(s) modelo(s) em aplicativos de teste no aplicativo real, se as restrições de tempo e orçamento permitirem.

Além disso, a avaliação também avalia outros resultados de mineração de dados gerados. Os resultados da mineração de dados envolvem modelos que estão necessariamente relacionados aos objetivos de negócios originais e todas as outras descobertas que não estão necessariamente relacionadas aos objetivos de negócios originais, mas também podem revelar desafios adicionais, informações ou dicas para direções futuras.

Saídas

Avaliação dos resultados da mineração de dados em relação aos critérios de sucesso do negócio Resumir os resultados da avaliação em termos de critérios de sucesso do negócio, incluindo uma declaração final sobre se o projeto já atende aos objetivos iniciais do negócio.

Modelos aprovados

Depois de avaliar os modelos com relação aos critérios de sucesso do negócio, os modelos gerados que atendem aos critérios selecionados tornam-se os modelos aprovados.

5.2 Processo de revisão

Tarefa

Processo de revisão Neste ponto, os modelos resultantes parecem satisfatórios e atendem às necessidades do negócio. Agora é apropriado fazer uma revisão mais completa do engajamento de mineração de dados para determinar se há algum fator ou tarefa importante que tenha sido negligenciado de alguma forma. Esta revisão também abrange questões de garantia de qualidade, por exemplo: Construímos o modelo corretamente? Usamos apenas os atributos que podemos usar e que estão disponíveis para análises futuras?

Saída

Revisão do processo

Resuma a revisão do processo e destaque as atividades que foram perdidas e aquelas que devem ser repetidas.

5.3 Determinar os próximos passos

Tarefa

Determinar as próximas etapas Dependendo dos resultados da avaliação e da revisão do processo, a equipe do projeto decide como proceder. A equipe decide se conclui esse projeto e segue para a implantação, inicia outras iterações ou configura novos projetos de mineração de dados. Esta tarefa inclui análises de recursos e orçamento remanescentes, que podem influenciar as decisões.

Saídas

Lista de ações possíveis

Liste as possíveis ações futuras, juntamente com as razões a favor e contra cada opção.

Decisão

Descreva a decisão sobre como proceder, juntamente com a justificativa.

6 Implantação

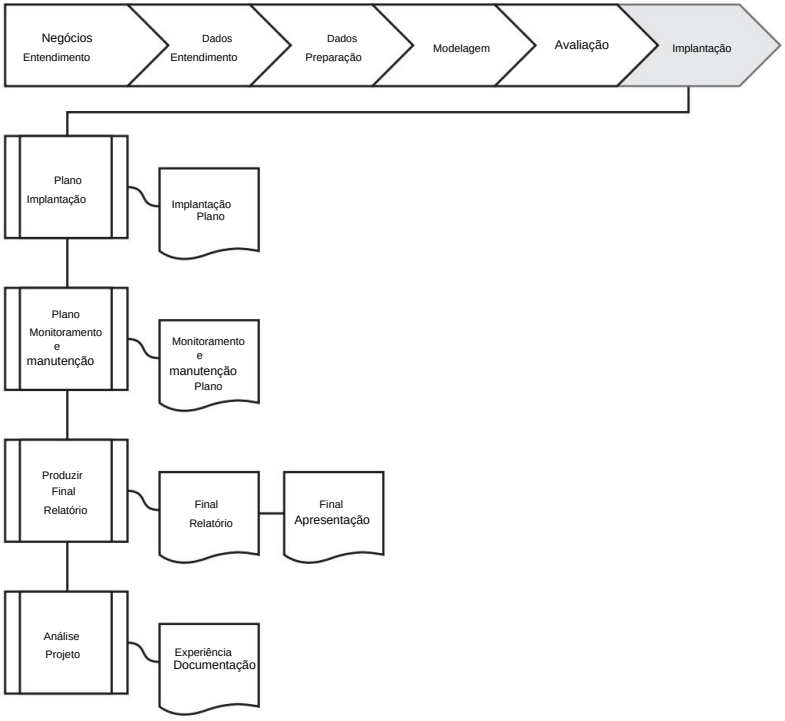

Figura 9: Implantação

6.1 Implantação do plano

Tarefa **Planejar a implantação** Esta tarefa obtém os resultados da avaliação e determina uma estratégia para a implantação. Se um procedimento geral foi identificado para criar o(s) modelo(s) relevante(s), esse procedimento é documentado aqui para implantação posterior.

Saída **Plano de implantação** Resuma a estratégia de implantação, incluindo as etapas necessárias e como executá-las.

6.2 Monitoramento e manutenção do plano

Tarefa

Monitoramento e manutenção do plano O monitoramento e a manutenção são questões importantes se o resultado da mineração de dados se tornar parte do dia-a-dia dos negócios e de seu ambiente. A preparação cuidadosa de uma estratégia de manutenção ajuda a evitar períodos desnecessariamente longos de uso incorreto dos resultados da mineração de dados. Para monitorar a implantação do(s) resultado(s) da mineração de dados, o projeto precisa de um plano de processo de monitoramento detalhado. Este plano leva em consideração o tipo específico de implantação.

Saída

Plano de monitoramento e manutenção

Resuma a estratégia de monitoramento e manutenção, incluindo as etapas necessárias e como executá-las.

6.3 Produzir relatório final

Tarefa

Produzir relatório final

Ao final do projeto, a equipe do projeto elabora um relatório final. Dependendo do plano de implantação, este relatório pode ser apenas um resumo do projeto e suas experiências (caso ainda não tenham sido documentadas como uma atividade em andamento) ou pode ser uma apresentação final e abrangente do(s) resultado(s) da mineração de dados.

Saídas

Relatório final Este é o relatório escrito final do trabalho de mineração de dados. Inclui todas as entregas anteriores, resumindo e organizando os resultados.

Apresentação final

Frequentemente também haverá uma reunião na conclusão do projeto em que os resultados são apresentados para o cliente.

6.4 Revisão do projeto

Tarefa

Reveja o projeto

Avalie o que deu certo e o que deu errado, o que foi bem feito e o que precisa ser melhorado.

Saída

Documentação da experiência

Resuma a experiência importante adquirida durante o projeto. Por exemplo, armadilhas, abordagens enganosas ou dicas para selecionar as técnicas de mineração de dados mais adequadas em situações semelhantes podem fazer parte desta documentação. Em projetos ideais, a documentação de experiência também cobre quaisquer relatórios que tenham sido escritos por membros individuais do projeto durante as fases anteriores do projeto.

III O guia do usuário do CRISP-DM

1 Compreensão do negócio

1.1 Determinar os objetivos de negócios

Tarefa

Determinar os objetivos de negócios O primeiro objetivo do analista é entender completamente, de uma perspectiva de negócios, o que o cliente realmente deseja realizar. Freqüentemente, o cliente tem muitos objetivos e restrições conflitantes que devem ser devidamente equilibrados. O objetivo do analista é descobrir fatores importantes no início do projeto que podem influenciar o resultado final. Uma consequência provável de negligenciar esta etapa seria gastar muito esforço produzindo as respostas corretas para as perguntas erradas.

Saída

Histórico

Reúna as informações conhecidas sobre a situação comercial da organização no início do projeto. Esses detalhes servem não apenas para identificar mais de perto os objetivos de negócios a serem alcançados, mas também para identificar os recursos, humanos e materiais, que podem ser usados ou necessários no decorrer do projeto.

Atividades

Organização ÿ

Desenvolva organogramas identificando divisões, departamentos e grupos de projetos. O gráfico deve também identifique os nomes e responsabilidades dos gerentes ÿ Identificar as pessoas-chave no negócio e suas funções ÿ Identificar um patrocinador interno (patrocinador financeiro e usuário principal/especialista no domínio) ÿ Indicar se há um comitê de direção e listar os membros ÿ Identificar as unidades de negócios que são afetados pelo projeto de mineração de dados (por exemplo, Marketing, Vendas, Finanças)

Área problemática

ÿ Identifique a área do problema (por exemplo, marketing, atendimento ao cliente, desenvolvimento de negócios, etc.) ÿ Descreva o problema em termos gerais ÿ Verifique o status atual do projeto (por exemplo, verifique se já está claro na unidade de negócios que um projeto de mineração deve ser realizado, ou se a mineração de dados precisa ser promovida como uma tecnologia chave no negócio)

ÿ Esclarecer os pré-requisitos do projeto (por exemplo, qual é a motivação do projeto? O negócio já usa mineração de dados?)

ÿ Se necessário, preparar apresentações e apresentar mineração de dados para o negócio

ÿ Identificar grupos-alvo para o resultado do projeto (por exemplo, devemos entregar um relatório para a alta administração ou um sistema operacional para ser usado por usuários finais ingênuos?)
ÿ Identificar as necessidades e expectativas dos usuários

solução atual

ÿ Descreva qualquer solução atualmente usada para resolver o problema ÿ Descreva as vantagens e desvantagens da solução atual e o nível em que ela é aceita pelos usuários

Saída

Objetivos de negócios

Descreva o objetivo principal do cliente, de uma perspectiva de negócios. Além do objetivo comercial principal, normalmente há um grande número de questões comerciais relacionadas que o cliente gostaria de abordar. Por exemplo, o principal objetivo de negócios pode ser manter os clientes atuais prevendo quando eles estão propensos a mudar para um concorrente, enquanto um objetivo de negócios secundário pode ser determinar se taxas mais baixas afetam apenas um segmento específico de clientes.

Atividades

ÿ Descreva informalmente o problema a ser resolvido ÿ Especifique todas as questões de negócios com a maior precisão possível ÿ Especifique quaisquer outros requisitos de negócios (por exemplo, a empresa não deseja perder nenhum cliente) ÿ Especifique os benefícios esperados em termos de negócios

Cuidado!

ÿ Cuidado ao estabelecer metas inatingíveis — torne-as o mais realistas possível.

Saída

Critérios de sucesso do negócio

Descreva os critérios para um resultado bem-sucedido ou útil para o projeto do ponto de vista do negócio. Isso pode ser bastante específico e facilmente mensurável, como a redução da rotatividade de clientes a um determinado nível, ou geral e subjetivo, como "fornecer informações úteis sobre os relacionamentos". Neste último caso, certifique-se de indicar quem faria o julgamento subjetivo.

Atividades

ÿ Especifique os critérios de sucesso comercial (por exemplo, melhorar a taxa de resposta em uma campanha de mala direta em 10 por cento e taxa de inscrição em 20 por cento)
ÿ Identifique quem avalia os critérios de sucesso

Lembrar! Cada um dos critérios de sucesso deve estar relacionado a pelo menos um dos objetivos de negócios especificados.

Boa ideia! Antes de iniciar a avaliação da situação, você pode analisar experiências anteriores desse problema — internamente, usando o CRISP-DM, ou externamente, usando soluções pré-empacotadas.

1.2 Avalie a situação

Tarefa **Avalie a situação**

Essa tarefa envolve a descoberta de fatos mais detalhados sobre todos os recursos, restrições, suposições e outros fatores que devem ser considerados na determinação do objetivo da análise de dados e no desenvolvimento do plano do projeto.

Saída **Inventário de recursos**

Liste os recursos disponíveis para o projeto, incluindo pessoal (especialistas em negócios e dados, suporte técnico, especialistas em mineração de dados), dados (extratos fixos, acesso a dados operacionais ou armazenados ao vivo), recursos de computação (plataformas de hardware) e software (ferramentas de mineração de dados, outro software relevante).

Atividades **Recursos de hardware**

ÿ Identifique o hardware base ÿ
Estabeleça a disponibilidade do hardware base para o projeto de mineração de dados ÿ Verifique
se o cronograma de manutenção do hardware está em conflito com a disponibilidade do hardware para o
projeto de mineração de dados
ÿ Identifique o hardware disponível para a ferramenta de mineração de dados a ser usada (se a ferramenta for conhecida nesta fase)

Fontes de dados e conhecimento ÿ
Identificar fontes de dados ÿ
Identificar tipos de fontes de dados (fontes online, especialistas, documentação escrita, etc.) ÿ Identificar
fontes de conhecimento ÿ Identificar
tipos de fontes de conhecimento (fontes online, especialistas, documentação escrita, etc.) ÿ Verifique as ferramentas
e técnicas disponíveis ÿ Descreva o conhecimento
prévio relevante (informalmente ou formalmente)

fontes de pessoal

ÿ Identifique o patrocinador do projeto (se for diferente do patrocinador interno conforme na Seção
1.1.1) ÿ Identifique o administrador do sistema, o administrador do banco de dados e a equipe de suporte técnico para mais perguntas ÿ
Identifique analistas de mercado, especialistas em mineração de dados e estatísticos e verifique sua disponibilidade ÿ
Verifique disponibilidade de especialistas de domínio para fases posteriores

Lembrar!

Lembre-se de que o projeto pode precisar de equipe técnica em momentos estranhos ao longo do projeto, por exemplo, durante a transformação de dados.

Saída

Requisitos, premissas e restrições Liste todos os requisitos do projeto, incluindo cronograma de conclusão, compreensibilidade e qualidade dos resultados e segurança, bem como questões legais. Como parte desta saída, certifique-se de que você tem permissão para usar os dados.

Liste as suposições feitas pelo projeto. Podem ser suposições sobre os dados, que podem ser verificadas durante a mineração de dados, mas também podem incluir suposições não verificáveis relacionadas ao projeto. É particularmente importante listar estes últimos se eles afetarem a validade dos resultados.

Liste as restrições feitas no projeto. Essas restrições podem envolver a falta de recursos para realizar algumas das tarefas do projeto no tempo necessário, ou pode haver restrições legais ou éticas sobre o uso dos dados ou a solução necessária para realizar a tarefa de mineração de dados.

Atividades

Requisitos ÿ
Especificar o perfil do grupo-alvo ÿ
Capturar todos os requisitos de agendamento ÿ
Capturar requisitos de compreensibilidade, precisão, capacidade de implantação, capacidade de manutenção e repetibilidade do projeto de mineração de dados e o(s) modelo(s) resultante(s)
ÿ Capturar requisitos de segurança, restrições legais, privacidade, relatórios e cronograma do projeto

Suposições ÿ
Esclarecer todas as suposições (incluindo as implícitas) e torná-las explícitas (por exemplo, para abordar a questão comercial, é necessário um número mínimo de clientes com idade acima de 50 anos)
ÿ Liste as suposições sobre a qualidade dos dados (por exemplo, precisão, disponibilidade) ÿ Liste as suposições sobre fatores externos (por exemplo, questões econômicas, produtos competitivos, avanços técnicos) ÿ Esclareça as suposições que levam a qualquer uma das estimativas (por exemplo, o preço de uma ferramenta específica é suposto ser inferior a US$ 1.000)
ÿ Liste todas as suposições sobre se é necessário entender e descrever ou explicar o modelo (por exemplo, como o modelo e os resultados devem ser apresentados à alta administração/patrocinador)

Restrições

ÿ Verifique as restrições gerais (por exemplo, questões legais, orçamento, prazos e recursos) ÿ
Verifique os direitos de acesso às fontes de dados (por exemplo, restrições de acesso, senha necessária)
ÿ Verifique a acessibilidade técnica dos dados (sistemas operacionais, sistema de gerenciamento de dados, arquivo ou banco de dados formato) ÿ Verifique se o conhecimento relevante está
acessível ÿ Verifique as restrições orçamentárias (custos fixos, custos de implementação, etc.)

Lembrar! A lista de suposições também inclui suposições no início do projeto, ou seja, qual foi o ponto de partida do projeto.

Saída

Riscos e contingências Liste os riscos, ou seja, os eventos que podem ocorrer, impactando cronograma, custo ou resultado. Liste os planos de contingência correspondentes: que ação será tomada para evitar ou minimizar o impacto ou recuperar da ocorrência dos riscos previstos.

Atividades

Identificar os riscos ÿ Identificar os riscos do negócio (por exemplo, o concorrente apresenta melhores resultados primeiro) ÿ Identificar os riscos organizacionais (por exemplo, o departamento que solicita o projeto não tem financiamento para o projeto) ÿ Identificar os riscos financeiros (por exemplo, financiamento adicional depende da mineração de dados inicial resultados) ÿ
Identificar riscos técnicos ÿ Identificar riscos que dependem de dados e fontes de dados (por exemplo, baixa qualidade e cobertura)

Desenvolver planos de contingência ÿ Determinar as condições sob as quais cada risco pode ocorrer ÿ Desenvolver planos de contingência

Saída

Terminologia

Compile um glossário de terminologia relevante para o projeto. Isso deve incluir pelo menos dois componentes: (1) Um glossário de terminologia de negócios relevante, que faz parte do entendimento de negócios disponível para o projeto
(2) Um glossário de terminologia de mineração de dados, ilustrado com exemplos relevantes para o negócio problema em questão

Atividades

ÿ Verifique a disponibilidade prévia de glossários; caso contrário, comece a redigir glossários ÿ
Converse com especialistas de domínio para entender sua terminologia ÿ
Familiarize-se com a terminologia comercial

Saída

Custos e benefícios

Prepare uma análise de custo-benefício para o projeto, comparando os custos do projeto com o potencial benefícios para o negócio se for bem sucedido

Atividades

ÿ Estimar custos para coleta de dados
ÿ Estimar custos de desenvolvimento e implementação de uma solução ÿ Identificar benefícios (por exemplo, maior satisfação do cliente, ROI e aumento na receita)
ÿ Estimar custos operacionais

Boa ideia!

A comparação deve ser o mais específica possível, pois isso permite que um melhor caso de negócios seja feito.

Cuidado!

Lembre-se de identificar os custos ocultos, como extração e preparação repetidas de dados, alterações nos fluxos de trabalho e tempo necessário para treinamento.

1.3 Determinar metas de mineração de dados

Tarefa

Determinar metas de mineração de dados Uma meta de negócios declara os objetivos na terminologia de negócios; uma meta de mineração de dados declara os objetivos do projeto em termos técnicos. Por exemplo, a meta de negócios pode ser “Aumentar as vendas do catálogo para clientes existentes”, enquanto uma meta de mineração de dados pode ser “Prever quantos widgets um cliente comprará, considerando suas compras nos últimos três anos, informações demográficas relevantes e o preço do item”.

Saída

Metas de mineração de dados Descreva as saídas pretendidas do projeto que permitem a realização dos objetivos de negócios. Observe que essas são normalmente saídas técnicas.

Atividades

ÿ Traduzir as questões de negócios para metas de mineração de dados (por exemplo, uma campanha de marketing requer segmentação de clientes para decidir quem abordar nessa campanha; o nível/tamanho dos segmentos deve ser especificado).

ÿ Especifique o tipo de problema de mineração de dados (por exemplo, classificação, descrição, predição e agrupamento). Para mais detalhes sobre os tipos de problemas de mineração de dados, consulte o Apêndice 2.

Boa ideia! Pode ser sábio redefinir o problema. Por exemplo, modelar a retenção de produtos em vez da retenção de clientes ao direcionar a retenção de clientes fornece resultados tarde demais para afetar o resultado.

Saída **Critérios de sucesso da mineração de dados** Defina os critérios para um resultado bem-sucedido do projeto em termos técnicos, por exemplo, um certo nível de precisão preditiva ou um perfil de propensão a comprar com um determinado grau de "elevação". Tal como acontece com os critérios de sucesso empresarial, pode ser necessário descrevê-los em termos subjetivos, caso em que a pessoa ou pessoas que fazem o julgamento subjetivo devem ser identificadas.

Atividades

- Especificar critérios para avaliação do modelo (por exemplo, precisão, desempenho e complexidade do modelo)
- Definir benchmarks para critérios de avaliação
- Especificar critérios que abordam critérios de avaliação subjetivos (por exemplo, capacidade de explicar o modelo e dados e insights de marketing fornecidos pelo modelo)

Cuidado! Lembre-se de que os critérios de sucesso da mineração de dados são diferentes dos critérios de sucesso do negócio definidos anteriormente.

Lembre-se de que é aconselhável planejar a implantação desde o início do projeto.

1.4 Produzir plano de projeto

Tarefa **Produzir plano de projeto**

Descreva o plano pretendido para atingir as metas de mineração de dados e, assim, atingir as metas de negócios.

Saída **Plano do projeto** Liste as etapas a serem executadas no projeto, juntamente com sua duração, recursos necessários, entradas, saídas e dependências. Sempre que possível, torne explícitas as iterações em larga escala no processo de mineração de dados — por exemplo, repetições das fases de modelagem e avaliação. Como parte do plano do projeto, também é importante analisar as dependências entre o cronograma e os riscos. Marque os resultados dessas análises explicitamente no plano do projeto, de preferência com ações e recomendações de ações caso os riscos se manifestem.

Embora esta seja a única tarefa em que o plano do projeto é nomeado diretamente, ele deve ser consultado continuamente e revisado ao longo do projeto. O plano do projeto deve ser consultado no mínimo sempre que uma nova tarefa é iniciada ou uma nova iteração de uma tarefa ou atividade é iniciada.

Atividades

ÿ Definir o plano de processo inicial e discutir a viabilidade com todo o pessoal envolvido ÿ Combinar todos os objetivos identificados e técnicas selecionadas em um procedimento coerente que resolva as questões de negócios e atenda aos critérios de sucesso do negócio

ÿ Estimar o esforço e os recursos necessários para alcançar e implantar a solução. (É útil considerar a experiência de outras pessoas ao estimar prazos para projetos de mineração de dados. Por exemplo, muitas vezes é postulado que 50-70 por cento do tempo e esforço em um projeto de mineração de dados é usado na fase de preparação de dados e 20-30 por cento na fase de compreensão de dados, enquanto apenas 10-20 por cento são gastos em cada das Fases de Modelagem, Avaliação e Entendimento do Negócio e 5-10 por cento na Fase de Implantação.)

ÿ Identificar etapas críticas ÿ Marcar pontos de decisão ÿ Marcar pontos de revisão ÿ Identificar as principais iterações

Saída

Avaliação inicial de ferramentas e técnicas Ao final da primeira fase, a equipe do projeto realiza uma avaliação inicial de ferramentas e técnicas. Aqui, é importante selecionar uma ferramenta de mineração de dados que suporte vários métodos para diferentes etapas do processo, pois a seleção de ferramentas e técnicas pode influenciar todo o projeto.

Atividades

ÿ Crie uma lista de critérios de seleção para ferramentas e técnicas (ou use uma existente, se disponível) ÿ Escolha ferramentas e técnicas potenciais ÿ Avalie a adequação das técnicas ÿ Revise e priorize as técnicas aplicáveis de acordo com a avaliação de soluções alternativas

2 Entendimento de dados

2.1 Coletar dados iniciais

Tarefa

Coletar dados iniciais

Adquira os dados (ou acesso aos dados) listados nos recursos do projeto. Essa coleta inicial inclui o carregamento de dados, se necessário para a compreensão dos dados. Por exemplo, se você pretende usar uma ferramenta específica para compreensão de dados, é lógico carregar seus dados nessa ferramenta.

Saída

Relatório inicial de coleta de dados

Descreva todos os vários dados usados para o projeto e inclua quaisquer requisitos de seleção para dados mais detalhados. O relatório de coleta de dados também deve definir se alguns atributos são relativamente mais importantes do que outros.

Lembre-se de que qualquer avaliação da qualidade dos dados deve ser feita não apenas das fontes de dados individuais, mas também de quaisquer dados resultantes da fusão de fontes de dados. Devido a inconsistências entre as fontes, os dados mesclados podem apresentar problemas que não existem nas fontes de dados individuais.

Atividades

Planejamento de requisitos de dados

ÿ Planejar quais informações são necessárias (por exemplo, apenas para determinados atributos ou informações adicionais específicas) ÿ Verificar se todas as informações necessárias (para resolver os objetivos de mineração de dados) estão realmente disponíveis

Critério de seleção

ÿ Especifique os critérios de seleção (por exemplo, quais atributos são necessários para os objetivos de mineração de dados especificados? Quais atributos foram identificados como irrelevantes? Quantos atributos podemos manipular com as técnicas escolhidas?) ÿ Selecione tabelas/arquivos de interesse ÿ Selecione dados dentro de uma tabela/arquivo ÿ Pense em quanto tempo um histórico deve ser usado (por exemplo, mesmo se 18 meses de dados estiverem disponíveis, apenas 12 meses podem ser necessários para o exercício)

Cuidado!

Esteja ciente de que os dados coletados de diferentes fontes podem dar origem a problemas de qualidade quando mesclados (por exemplo, arquivos de endereços mesclados com um banco de dados de clientes podem apresentar inconsistências de formato, invalidez de dados, etc.).

Inserção de dados

ÿ Se os dados contiverem entradas de texto livre, precisamos codificá-los para modelagem ou queremos agrupá-los entradas específicas?

ÿ Como os atributos que faltam podem ser adquiridos?

ÿ Qual a melhor forma de extrair os dados?

Boa ideia! Lembre-se de que algum conhecimento sobre os dados pode estar disponível em fontes não eletrônicas (por exemplo, de pessoas, texto impresso, etc.).

Lembre-se que pode ser necessário pré-processar os dados (dados de séries temporais, médias ponderadas, etc.).

2.2 Descrever dados

Tarefa **Descrever dados**

Examine as propriedades “brutas” dos dados adquiridos e relate os resultados.

Saída **Relatório de descrição de dados** Descreva os dados que foram adquiridos, incluindo o formato dos dados, a quantidade de dados (por exemplo, o número de registros e campos dentro de cada tabela), as identidades dos campos e quaisquer outros características de superfície que foram descobertas.

Atividades

Análise volumétrica de dados

ÿ Identificar dados e método de captura
ÿ Acessar fontes de dados
ÿ Use análises estatísticas, se apropriado ÿ
Tabelas de relatório e suas relações ÿ
Verifique o volume de dados, número de múltiplos, complexidade
ÿ Observe se os dados contêm entradas de texto livre

Tipos e valores de atributos ÿ
Verifique a acessibilidade e disponibilidade de atributos ÿ
Verifique os tipos de atributos (numéricos, simbólicos, taxonomia, etc.) ÿ
Verifique os intervalos de valores de
atributos ÿ Analise as correlações de
atributos ÿ Entenda o significado de cada atributo e valor de atributo em termos comerciais ÿ Para
cada atributo, calcular estatísticas básicas (por exemplo, calcular distribuição, média, máximo, mínimo, padrão desvio, variância, modo, assimetria, etc.)
ÿ Analisar estatísticas básicas e relacionar os resultados com seu significado em termos de negócios
ÿ Decidir se o atributo é relevante para a meta específica de mineração de dados

ÿ Determine se o significado do atributo é usado de forma consistente ÿ
Entreviste especialistas do domínio para obter sua opinião sobre a relevância do atributo ÿ Decida
se é necessário equilibrar os dados (com base nas técnicas de modelagem a serem usadas)

Chaves

ÿ Analisar relacionamentos de chave
ÿ Verificar a quantidade de sobreposições de valores de atributos de chave nas tabelas

Revisar suposições/metas ÿ
Atualizar lista de suposições, se necessário

2.3 Explorar dados

Tarefa

Explorar dados

Esta tarefa aborda as questões de mineração de dados que podem ser abordadas usando técnicas de consulta, visualização e geração de relatórios. Essas análises podem abordar diretamente os objetivos da mineração de dados. No entanto, eles também podem contribuir ou refinar a descrição dos dados e os relatórios de qualidade e alimentar a transformação e outras etapas de preparação de dados necessárias antes que uma análise mais aprofundada possa ocorrer.

Saída

Relatório de exploração de dados Descreva os resultados desta tarefa, incluindo as primeiras descobertas ou hipóteses iniciais e seu impacto no restante do projeto. O relatório também pode incluir gráficos e plotagens que indicam as características dos dados ou apontam para subconjuntos de dados interessantes que merecem um exame mais aprofundado.

Atividades

Exploração de dados

ÿ Analisar propriedades de atributos interessantes em detalhes (por exemplo, estatísticas básicas, subpopulações interessantes) ÿ Identificar características de subpopulações

Formule suposições para análise futura ÿ
Considere e avalie informações e descobertas no relatório de descrição de dados ÿ Formule uma hipótese e identifique ações ÿ Transforme a hipótese em uma meta de mineração de dados, se possível ÿ Esclareça as metas de mineração de dados ou torne-as mais precisas. Uma busca “cega” não é necessariamente inútil, mas uma busca mais direcionada aos objetivos de negócios é preferível. ÿ Realizar análise básica para verificar a hipótese

2.4 Verifique a qualidade dos dados

Tarefa

Verifique a qualidade dos dados Examine a qualidade dos dados, abordando questões como: Os dados estão completos (abrangem todos os casos necessários)? Está correto ou contém erros? Se houver erros, quão comuns eles são? Há valores ausentes nos dados? Em caso afirmativo, como são representados, onde ocorrem e quão comuns são?

Saída

Relatório de qualidade dos dados Liste os resultados da verificação da qualidade dos dados; se houver problemas de qualidade, liste possíveis soluções.

Atividades

ÿ Identificar valores especiais e catalogar seu significado

Revise as chaves, atributos ÿ

Verifique a cobertura (por exemplo, se todos os valores possíveis estão representados) ÿ Verifique as chaves ÿ

Verifique se os significados dos atributos e valores contidos se encaixam ÿ Identifique atributos ausentes e campos em branco ÿ Estabeleça o significado dos dados ausentes ÿ Verifique os atributos com valores diferentes que têm significados semelhantes (por exemplo, baixo teor de gordura, dieta) ÿ Verifique a ortografia e o formato dos valores (por exemplo, mesmo valor, mas às vezes começando com uma letra minúscula, às vezes com uma letra maiúscula)

ÿ Verifique se há desvios e decida se um desvio é "ruído" ou pode indicar um fenômeno interessante ÿ Verifique a plausibilidade dos valores (por exemplo, todos os campos com os mesmos ou quase os mesmos valores)

Boa ideia!

Revise todos os atributos que dão respostas que conflitam com o senso comum (por exemplo, adolescentes com altos níveis de renda).

Use gráficos de visualização, histogramas, etc. para revelar inconsistências nos dados.

Qualidade de dados em arquivos simples ÿ Se os dados forem armazenados em arquivos simples, verifique qual delimitador é usado e se ele é usado consistentemente dentro de todos os atributos

ÿ Se os dados estiverem armazenados em arquivos simples, verifique o número de campos em cada registro para ver se eles coincidem

Ruído e inconsistências entre as fontes

ÿ Verifique consistências e redundâncias entre diferentes fontes

ÿ Planejar para lidar com o ruído ÿ

Detectar o tipo de ruído e quais atributos são afetados

Boa ideia!

Recorde-se que poderá ser necessário excluir alguns dados uma vez que não apresentam comportamento nem positivo nem negativo (por exemplo, para verificar o comportamento de crédito dos clientes, excluir todos os que nunca tomaram empréstimo, não financiam habitação própria, aqueles cuja hipoteca é quase maturidade, etc.).

Revise se as suposições são válidas ou não, dadas as informações atuais sobre dados e conhecimento do negócio.

3 Preparação de dados

Saída conjunto de dados

Estes são os conjuntos de dados produzidos pela fase de preparação de dados, usados para modelagem ou para o trabalho de análise principal do projeto.

Saída **Descrição do conjunto de dados** Esta é a descrição do(s) conjunto(s) de dados usados para a modelagem ou para o trabalho de análise principal do projeto.

3.1 Selecione os dados

Tarefa **Selecionar dados**

Decidir sobre os dados a serem usados para análise. Os critérios incluem relevância para as metas de mineração de dados, qualidade e restrições técnicas, como limites de volume ou tipos de dados.

Saída **Justificativa da inclusão/exclusão** Liste os dados a serem utilizados/excluídos e os motivos dessas decisões.

Atividades

ÿ Colete dados adicionais apropriados (de diferentes fontes - internas e externas) ÿ Realize testes de significância e correlação para decidir se os campos devem ser incluídos ÿ Reconsidere os critérios de seleção de dados (consulte a Tarefa 2.1) à luz das experiências de qualidade de dados e dados exploração (ou seja, pode desejar incluir/excluir outros conjuntos de dados) ÿ Reconsidere os critérios de seleção de dados (consulte a tarefa 2.1) à luz da experiência de modelagem (ou seja, modelo avaliação pode mostrar que outros conjuntos de dados são necessários)

ÿ Selecione diferentes subconjuntos de dados (por exemplo, atributos diferentes, apenas dados que atendam a certas condições)

ÿ Considere o uso de técnicas de amostragem (por exemplo, uma solução rápida pode envolver a divisão de conjuntos de dados de teste e treinamento ou a redução do tamanho do conjunto de dados de teste, se a ferramenta não puder lidar com o conjunto de dados completo. Também pode ser útil ter amostras ponderadas para fornecer diferentes importância para diferentes atributos ou valores diferentes do mesmo

atributo.) ÿ Documente a justificativa para inclusão/exclusão ÿ

Verifique as técnicas disponíveis para amostragem de dados

Boa ideia!

Com base nos critérios de seleção de dados, decida se um ou mais atributos são mais importantes do que outros e pondere os atributos de acordo. Decida, com base no contexto (ou seja, aplicativo, ferramenta, etc.), como lidar com a ponderação.

3.2 Limpar dados

Tarefa **Limpar dados**

Aumente a qualidade dos dados para o nível exigido pelas técnicas de análise selecionadas. Isso pode envolver a seleção de subconjuntos limpos dos dados, a inserção de padrões adequados ou técnicas mais ambiciosas, como a estimativa de dados ausentes por modelagem.

Saída **Relatório de limpeza de dados** Descreva as decisões e ações que foram tomadas para resolver os problemas de qualidade de dados relatados durante a tarefa Verificar qualidade de dados. Se os dados forem usados no exercício de mineração de dados, o relatório deve abordar questões pendentes de qualidade de dados e o possível efeito que isso poderia ter nos resultados.

Atividades

ÿ Reconsidere como lidar com qualquer tipo de ruído observado

ÿ Corrija, remova ou ignore o ruído

ÿ Decida como lidar com valores especiais e seu significado. A área de valores especiais pode dar origem a muitos resultados estranhos e devem ser cuidadosamente examinados. Exemplos de valores especiais podem surgir ao obter resultados de uma pesquisa em que algumas perguntas não foram feitas ou não foram respondidas. Isso pode resultar em um valor de 99 para dados desconhecidos. Por exemplo, 99 para estado civil ou afiliação política. Valores especiais também podem surgir quando os dados são truncados - por exemplo, 00 para pessoas de 100 anos ou todos os carros com 100.000 km no hodômetro.

ÿ Reconsidere os critérios de seleção de dados (consulte a tarefa 2.1) à luz das experiências de limpeza de dados (ou seja, você pode deseja incluir/excluir outros conjuntos de dados).

Boa ideia!

Lembre-se de que alguns campos podem ser irrelevantes para os objetivos da mineração de dados e, portanto, o ruído nesses campos não tem significado. No entanto, se o ruído for ignorado por esses motivos, ele deve ser totalmente documentado, pois as circunstâncias podem mudar posteriormente.

3.3 Construir dados

Tarefa

Construir dados

Esta tarefa inclui operações construtivas de preparação de dados, como a produção de atributos derivados, novos registros completos ou valores transformados para atributos existentes.

Atividades

ÿ Verificar os mecanismos de construção disponíveis com a lista de ferramentas sugeridas para o projeto ÿ Decidir se é melhor fazer a construção dentro ou fora da ferramenta (ou seja, o que é mais eficiente, exato, repetível)

ÿ Reconsidere os critérios de seleção de dados (consulte a tarefa 2.1) à luz das experiências de construção de dados (ou seja, você pode desejar incluir/excluir outros conjuntos de dados)

Saída

Atributos derivados

Atributos derivados são novos atributos construídos a partir de um ou mais atributos existentes no mesmo registro. Um exemplo pode ser: área = comprimento * largura.

Por que precisamos construir atributos derivados durante uma investigação de mineração de dados? Não se deve pensar que apenas dados de bancos de dados ou outras fontes devem ser usados na construção de um modelo. Atributos derivados podem ser construídos porque:

ÿ O conhecimento prévio nos convence de que algum fato é importante e deve ser representado, embora não temos nenhum atributo atualmente para representá-lo ÿ O algoritmo de modelagem em uso lida apenas com certos tipos de dados - por exemplo, estamos usando linear regressão e suspeitamos que existem certas não linearidades que não serão incluídas no modelo

ÿ O resultado da fase de modelagem sugere que certos fatos não estão sendo cobertos

Atividades

Atributos derivados

ÿ Decida se algum atributo deve ser normalizado (por exemplo, ao usar um algoritmo de agrupamento com idade e renda, em certas moedas, a renda dominará)

ÿ Considere adicionar novas informações sobre a importância relevante dos atributos adicionando novos atributos (por exemplo, pesos de atributo, normalização ponderada)

ÿ Como os atributos ausentes podem ser construídos ou imputados? [Decidir o tipo de construção (por exemplo, agregada, média, indução).]
ÿ Adicionar novos atributos aos dados acessados

Boa ideia!

Antes de adicionar Atributos Derivados, tente determinar se e como eles facilitam o processo de modelagem ou facilitam o algoritmo de modelagem. Talvez "renda por pessoa" seja um atributo melhor/mais fácil de usar do que "renda por família". Não derivar atributos simplesmente para reduzir o número de atributos de entrada.

Outro tipo de atributo derivado é a transformação de atributo único, geralmente realizada para atender às necessidades das ferramentas de modelagem.

Atividades

Transformações de atributo único ÿ

Especifique as etapas de transformação necessárias em termos de instalações de transformação disponíveis (por exemplo, alterar um agrupamento de um atributo numérico)
ÿ Executar etapas de transformação

Boa ideia!

Transformações podem ser necessárias para alterar intervalos para campos simbólicos (por exemplo, idades para faixas etárias) ou campos simbólicos ("definitivamente sim", "sim", "não sei", "não") para valores numéricos. Ferramentas de modelagem ou algoritmos geralmente os exigem.

Saída

registros gerados

Registros gerados são registros completamente novos, que adicionam novo conhecimento ou representam novos dados que não são representados de outra forma (por exemplo, tendo segmentado os dados, pode ser útil gerar um registro para representar o membro protótipo de cada segmento para processamento posterior).

Atividades

Verifique as técnicas disponíveis, se necessário (por exemplo, mecanismos para construir protótipos para cada segmento de dados segmentados).

3.4 Integrar dados

Tarefa

Integrar dados

Estes são métodos para combinar informações de várias tabelas ou outras fontes de informação para criar novos registros ou valores.

Saída

Dados

mesclados Mesclar tabelas refere-se à junção de duas ou mais tabelas que possuem informações diferentes sobre os mesmos objetos.

Nesta fase, também pode ser aconselhável gerar novos registros. Também pode ser recomendado gerar valores agregados.

Agregação refere-se a operações em que novos valores são calculados resumindo informações de vários registros e/ou tabelas.

Atividades

ÿ Verifique se os recursos de integração são capazes de integrar as fontes de entrada conforme necessário ÿ Integre as fontes e armazene os resultados ÿ Reconsidere os critérios de seleção de dados (consulte a Tarefa 2.1) à luz das experiências de integração de dados (ou seja, você pode desejar incluir/excluir outros conjuntos de dados)

Boa ideia!

Lembre-se de que alguns conhecimentos podem estar contidos em formato não eletrônico.

3.5 Formatar dados

Tarefa

Dados de formato

As transformações de formatação referem-se principalmente a modificações sintáticas feitas nos dados que não alteram seu significado, mas podem ser exigidas pela ferramenta de modelagem.

Saída

dados reformatados

Algumas ferramentas têm requisitos na ordem dos atributos, como o primeiro campo sendo um identificador exclusivo para cada registro ou o último campo sendo o campo de resultado que o modelo deve prever.

Atividades

Reorganizando atributos

Algumas ferramentas têm requisitos na ordem dos atributos, como o primeiro campo sendo um identificador exclusivo para cada registro ou o último campo sendo o campo de resultado que o modelo deve prever.

Reordenar registros Pode

ser importante alterar a ordem dos registros no conjunto de dados. Talvez a ferramenta de modelagem exija que os registros sejam classificados de acordo com o valor do atributo de resultado.

dentro do valor reformatado

ÿ Estas são alterações puramente sintáticas feitas para satisfazer os requisitos da ferramenta de modelagem específica ÿ Reconsidere os critérios de seleção de dados (consulte a tarefa 2.1) à luz das experiências de limpeza de dados (ou seja, você pode querer incluir/excluir outros conjuntos de dados)

4 Modelagem

4.1 Selecione a técnica de modelagem

Tarefa

Selecione a técnica de modelagem

Como primeira etapa na modelagem, selecione a técnica de modelagem inicial real. Se várias técnicas forem aplicadas, execute esta tarefa separadamente para cada técnica.

Lembre-se de que nem todas as ferramentas e técnicas são aplicáveis a todas as tarefas. Para certos problemas, apenas algumas técnicas são apropriadas (consulte o Apêndice 2, onde as técnicas apropriadas para certos tipos de problemas de mineração de dados são discutidas com mais detalhes). “Requisitos políticos” e outras restrições limitam ainda mais as opções disponíveis para o engenheiro de mineração de dados. Pode ser que apenas uma ferramenta ou técnica esteja disponível para resolver o problema em questão – e que a ferramenta pode não ser absolutamente a melhor, do ponto de vista técnico.

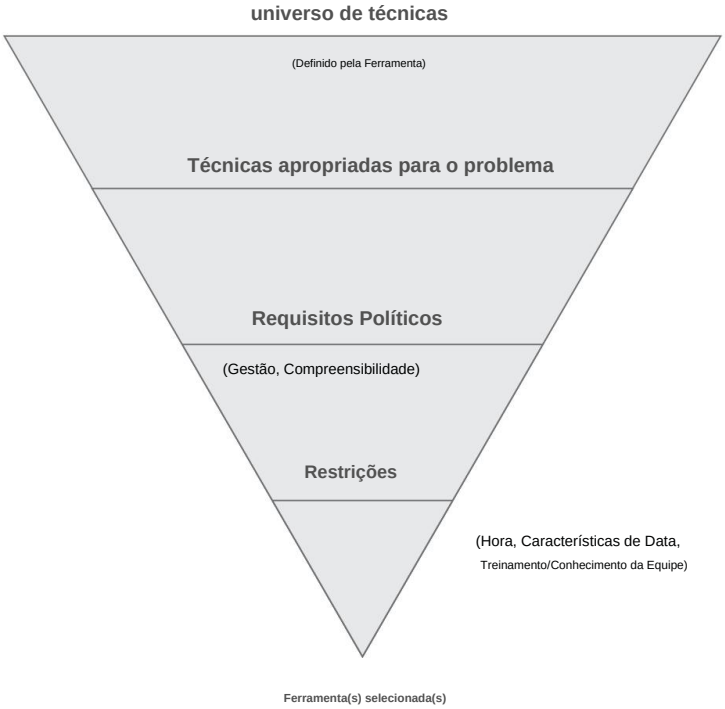

Figura 10:
universo de técnicas

Saída	**Técnica de modelagem** Registre a técnica de modelagem real que é usada.
Atividades	Decida a técnica apropriada para o exercício, tendo em mente a ferramenta selecionada.
Saída	**Suposições de** **modelagem** Muitas técnicas de modelagem fazem suposições específicas sobre os dados.
Atividades	ÿ Defina quaisquer suposições incorporadas feitas pela técnica sobre os dados (por exemplo, qualidade, formato, distribuição) ÿ Compare essas suposições com as do Relatório de descrição de dados ÿ Certifique-se de que essas suposições sejam válidas e volte para a Fase de preparação de dados, se necessário

4.2 Gerar projeto de teste **Gerar**

Tarefa **projeto de teste** Antes de construir um modelo, é necessário definir um procedimento para testar a qualidade e validade do modelo. Por exemplo, em tarefas de mineração de dados supervisionadas, como classificação, é comum usar taxas de erro como medidas de qualidade para modelos de mineração de dados. Portanto, o design do teste especifica que o conjunto de dados deve ser separado em conjuntos de treinamento e teste. O modelo é construído no conjunto de treinamento e sua qualidade estimada no conjunto de teste.

Saída **Projeto de teste** Descreva o plano pretendido para treinar, testar e avaliar os modelos. Um componente primário do plano é decidir como dividir o conjunto de dados disponível em dados de treinamento, dados de teste e conjuntos de teste de validação.

Atividades ÿ Verifique os designs de teste existentes para cada meta de mineração de dados separadamente ÿ Decida as etapas necessárias (número de iterações, número de dobras, etc.) ÿ Prepare os dados necessários para o teste

4.3 Construir modelo

Tarefa **Construir modelo**

Execute a ferramenta de modelagem no conjunto de dados preparado para criar um ou mais modelos.

Saída **Configurações de parâmetros** Com qualquer ferramenta de modelagem, geralmente há um grande número de parâmetros que podem ser ajustados. Liste os parâmetros e seus valores escolhidos, juntamente com a justificativa para a escolha.

Atividades ÿ Defina os parâmetros iniciais ÿ Documente os motivos para escolher esses valores

Saída **modelos**

Execute a ferramenta de modelagem no conjunto de dados preparado para criar um ou mais modelos.

Atividades ÿ Executar a técnica selecionada no conjunto de dados de entrada para produzir o modelo ÿ Resultados de mineração de dados pós-processamento (por exemplo, editar regras, exibir árvores)

Saída **Descrição do modelo**

Descreva o modelo resultante e avalie sua precisão esperada, robustez e possíveis deficiências.

Relatório sobre a interpretação dos modelos e eventuais dificuldades encontradas.

Atividades

ÿ Descreva quaisquer características do modelo atual que possam ser úteis para o futuro ÿ Registre as configurações de parâmetros usadas para produzir o modelo ÿ Forneça uma descrição detalhada do modelo e quaisquer recursos especiais ÿ Para modelos baseados em regras, liste as regras produzidas, mais quaisquer avaliação da precisão e cobertura por regra ou modelo geral

ÿ Para modelos opacos, liste todas as informações técnicas sobre o modelo (como a topologia da rede neural) e quaisquer descrições comportamentais produzidas pelo processo de modelagem (como precisão ou sensibilidade) ÿ Descreva o comportamento e a interpretação do modelo ÿ Declare as conclusões sobre os padrões nos dados (caso existam); às vezes o modelo revela fatos importantes sobre os dados sem um processo de avaliação separado (por exemplo, que a saída ou conclusão é duplicada em uma das entradas)

4.4 Avaliar o modelo

Tarefa **Avaliar modelo**

O modelo agora deve ser avaliado para garantir que atenda aos critérios de sucesso da mineração de dados e passe nos critérios de teste desejados. Esta é uma avaliação puramente técnica baseada no resultado das tarefas de modelagem.

Saída **Avaliação do modelo**

Resuma os resultados desta tarefa, liste as qualidades dos modelos gerados (por exemplo, em termos de precisão) e classifique sua qualidade em relação uns aos outros.

Atividades

ÿ Avalie os resultados em relação aos critérios de avaliação ÿ O resultado do teste de acordo com uma estratégia de teste (por exemplo: treinamento e teste, validação cruzada, bootstrap, etc.) ÿ Compare os resultados e a interpretação da avaliação ÿ Crie uma classificação dos resultados com relação aos critérios de sucesso e avaliação ÿ Selecione os melhores modelos

ÿ Interprete os resultados em termos de negócios (na medida do possível neste estágio) ÿ Obtenha comentários sobre modelos por especialistas em domínio ou dados ÿ Verifique a plausibilidade do modelo

ÿ Verifique o efeito na meta de mineração de
dados ÿ Verifique o modelo em relação à base de conhecimento fornecida para ver se a informação descoberta é nova e útil ÿ Verifique a
confiabilidade do resultado ÿ Analise
o potencial para implantação de cada resultado ÿ Se houver uma
descrição verbal do modelo gerado (por exemplo , via regras), avaliar as regras: São lógicas,
são viáveis, são muitos ou poucos, ofendem o bom senso?
ÿ Avaliar resultados
ÿ Obtenha informações sobre por que uma determinada técnica de modelagem e certas configurações de parâmetros levam a resultados bons/ruins

Boa ideia! "Tabelas de elevação" e "Tabelas de ganho" podem ser construídas para determinar o quão bem o modelo está prevendo.

Saída **Configurações de parâmetro**
revisadas De acordo com a avaliação do modelo, revise as configurações de parâmetro e ajuste-as para a próxima execução na tarefa Construir modelo. Repita a construção e a avaliação do modelo até encontrar o melhor modelo.

Atividades Ajuste os parâmetros para produzir modelos melhores.

5 Avaliação

As etapas de avaliação anteriores lidaram com fatores como a precisão e a generalidade do modelo. Esta etapa avalia o grau em que o modelo atende aos objetivos de negócios e procura determinar se há algum motivo comercial para que esse modelo seja deficiente. Ele compara os resultados com os critérios de avaliação definidos no início do projeto.

Uma boa maneira de definir as saídas totais de um projeto de mineração de dados é usar a equação:

RESULTADOS = MODELOS + CONCLUSÕES

Nesta equação, estamos definindo que a saída total do projeto de mineração de dados não é apenas os modelos (embora sejam, é claro, importantes), mas também as descobertas, que definimos como qualquer coisa (além do modelo) que é importante no cumprimento dos objetivos do negócio ou importante para levar a novas questões, linhas de abordagem ou efeitos colaterais (por exemplo, problemas de qualidade de dados descobertos pelo exercício de mineração de dados). Observação: embora o modelo esteja diretamente conectado às questões de negócios, as descobertas não precisam estar relacionadas a nenhuma questão ou objetivo, desde que sejam importantes para o iniciador do projeto.

5.1 Avaliar resultados

Tarefa **Avalie os resultados**

Esta etapa avalia o grau em que o modelo atende aos objetivos de negócios e procura determinar se há algum motivo comercial para que esse modelo seja deficiente. Outra opção é testar o(s) modelo(s) em aplicativos de teste no aplicativo real, se as restrições de tempo e orçamento permitirem.

Além disso, a avaliação também avalia outros resultados de mineração de dados gerados. Os resultados da mineração de dados abrangem modelos relacionados aos objetivos de negócios originais e todas as outras descobertas. Alguns estão relacionados aos objetivos de negócios originais, enquanto outros podem revelar desafios, informações ou dicas adicionais para direções futuras.

Saída **Avaliação dos resultados da mineração de dados em relação aos critérios de sucesso do negócio** Resumir os resultados da avaliação em termos de critérios de sucesso do negócio, incluindo uma declaração final relacionada a se o projeto já atende aos objetivos iniciais do negócio.

Atividades

ÿ Entenda os resultados da mineração de dados

ÿ Interprete os resultados em termos do aplicativo ÿ Verifique

o efeito para a meta de mineração de dados ÿ

Verifique o resultado da mineração de dados em relação à base de conhecimento fornecida para ver se as informações descobertas é novo e útil

ÿ Avaliar e avaliar os resultados com relação aos critérios de sucesso do negócio (ou seja, o projeto alcançou os Objetivos Comerciais originais)

ÿ Comparar os resultados e a interpretação da avaliação ÿ

Classificar os resultados em relação aos critérios de sucesso do

negócio ÿ Verificar o efeito do resultado na meta inicial da

aplicação ÿ Determinar se há novos objetivos de negócios a serem abordados posteriormente no projeto ou em novos projetos ÿ

Indicar recomendações para dados futuros projetos de mineração

Saída **modelos aprovados**

Após acessar os modelos com relação aos critérios de sucesso do negócio, selecione e aprove os modelos gerados que atendem aos critérios selecionados.

5.2 Processo de revisão

Tarefa **Processo de revisão** Neste ponto, o modelo resultante parece ser satisfatório e parece satisfazer as necessidades de negócios. Agora é apropriado fazer uma revisão mais completa do engajamento de mineração de dados para determinar se há algum fator ou tarefa importante que tenha sido negligenciado de alguma forma. Nesta fase do exercício de mineração de dados, a Revisão do Processo assume a forma de uma Revisão de Garantia de Qualidade.

Saída **Revisão do processo**

Resuma a revisão do processo e liste as atividades que foram perdidas e/ou devem ser repetidas.

Atividades

ÿ Fornecer uma visão geral do processo de mineração de dados usado. ÿ Analisar o processo de mineração de dados. Para cada etapa do processo, pergunte:

- ÿ Foi necessário?
- ÿ Foi executado de forma otimizada?
- ÿ De que forma poderia ser melhorado?

ÿ Identificar falhas ÿ Identificar etapas enganosas ÿ Identificar possíveis ações alternativas e/ou caminhos inesperados no processo ÿ Analisar os resultados da mineração de dados em relação aos critérios de sucesso do negócio

5.3 Determinar os próximos passos

Tarefa **Determinar as próximas etapas** Com base nos resultados da avaliação e na revisão do processo, a equipe do projeto decide como proceder. As decisões a serem tomadas incluem terminar este projeto e passar para a implantação, iniciar outras iterações ou configurar novos projetos de mineração de dados.

Saída **Lista de ações possíveis**

Liste outras ações possíveis junto com as razões a favor e contra cada opção.

Atividades

ÿ Analisar o potencial de implantação de cada resultado ÿ Estimar o potencial de melhoria do processo atual ÿ Verificar os recursos restantes para determinar se eles permitem iterações de processo adicionais (ou se recursos adicionais podem ser disponibilizados)

ÿ Recomendar continuações alternativas

ÿ Refinar plano de processo

Saída

Decisão

Descreva as decisões tomadas, juntamente com a justificativa para elas.

Atividades

ÿ Classifique as ações possíveis ÿ Selecione uma das ações possíveis

ÿ Razões do documento para a escolha

6 Implantação *6.1*

Planejar a implantação

Tarefa

Planejar a implantação Esta tarefa começa com os resultados da avaliação e termina com uma estratégia para a implantação do(s) resultado(s) da mineração de dados no negócio.

Saída

Plano de implantação

Resuma a estratégia de implantação, incluindo as etapas necessárias e como executá-las.

Atividades

ÿ Resumir resultados implantáveis ÿ Desenvolver e avaliar planos alternativos para implantação ÿ Decidir para cada resultado distinto de conhecimento ou informação ÿ Determinar como o conhecimento ou informação será propagado aos usuários ÿ Decidir como o uso do resultado será monitorado e seus benefícios medidos (quando aplicável) ÿ Decidir para cada modelo ou resultado de software implantável ÿ Estabelecer como o modelo ou resultado de software será implantado nos sistemas da organização ÿ Determinar como seu uso será monitorado e seus benefícios medidos (quando aplicável) ÿ Identificar possíveis problemas durante a implantação (armadilhas para ser evitado)

6.2 Monitoramento e manutenção do plano

Tarefa

Monitoramento e manutenção do plano O monitoramento e a manutenção são questões importantes se os resultados da mineração de dados se tornarem parte do dia-a-dia dos negócios e de seu ambiente. Uma preparação cuidadosa de uma estratégia de manutenção ajuda a evitar períodos desnecessariamente longos de uso incorreto dos resultados da mineração de dados. Para monitorar a implantação do(s) resultado(s) da mineração de dados, o projeto precisa de um plano detalhado de monitoramento e manutenção. Este plano leva em consideração o tipo específico de implantação.

Saída

Plano de monitoramento e manutenção

Resuma a estratégia de monitoramento e manutenção, incluindo as etapas necessárias e como executá-las.

Atividades

ÿ Verifique os aspectos dinâmicos (ou seja, o que pode mudar no ambiente?) ÿ Decida como a precisão será monitorada ÿ Determine quando o resultado ou modelo de mineração de dados não deve mais ser usado. Identifique os critérios (validade, limite de precisão, novos dados, mudança no domínio do aplicativo etc.) e o que deve acontecer se o modelo ou resultado não puder mais ser usado. (atualizar modelo, configurar novo projeto de mineração de dados, etc.). ÿ Os objetivos de negócios do uso do modelo mudarão com o tempo? Documente totalmente o problema inicial que o modelo estava tentando resolver.

ÿ Desenvolver plano de monitoramento e manutenção.

6.3 Produzir relatório final

Tarefa

Produzir relatório final

Ao final do projeto, a equipe do projeto elabora um relatório final. Dependendo do plano de implantação, este relatório pode ser apenas um resumo do projeto e sua experiência, ou uma apresentação final do(s) resultado(s) da mineração de dados.

Saída

Relatório final Ao final do projeto, haverá pelo menos um relatório final no qual todos os fios são reunidos. Além de identificar os resultados obtidos, o relatório também deve descrever o processo, mostrar quais custos foram incorridos, definir eventuais desvios do plano original, descrever planos de implementação e fazer recomendações para trabalhos futuros. O conteúdo real detalhado do relatório depende muito sobre o público-alvo.

Atividades

ÿ Identifique quais relatórios são necessários (apresentação de slides, resumo de gerenciamento, descobertas detalhadas, explicação dos modelos, etc.) ÿ
Analisar como as metas iniciais de mineração de dados foram atendidas ÿ
Identificar grupos-alvo para o relatório ÿ
Descrever a estrutura e o conteúdo do(s) relatório(s) ÿ
Selecionar descobertas a serem incluídas nos relatórios
ÿ Escreva um relatório

Saída

Apresentação final

Além do relatório final, pode ser necessário fazer uma apresentação final para resumir o projeto — talvez para o patrocinador administrativo, por exemplo. A apresentação normalmente contém um subconjunto da informação contida no relatório final, estruturada de forma diferente.

Atividades

ÿ Decidir sobre o grupo-alvo para a apresentação final e determinar se eles já receberam o relatório final
ÿ Selecione quais itens do relatório final devem ser incluídos na apresentação final

6.4 Revisão do projeto

Tarefa

Revisão do projeto Avalie o que deu certo e o que deu errado, o que foi bem feito e o que precisa ser melhorado.

Saída

Documentação da experiência

Resuma a experiência importante adquirida durante o projeto. Por exemplo, armadilhas, abordagens enganosas ou dicas para selecionar as técnicas de mineração de dados mais adequadas em situações semelhantes podem fazer parte desta documentação. Em projetos ideais, a documentação de experiência também cobre quaisquer relatórios que tenham sido escritos por membros individuais do projeto durante o projeto.

Atividades

ÿ Entreviste todas as pessoas importantes envolvidas no projeto e pergunte sobre sua experiência durante o projeto ÿ Se os usuários finais da empresa trabalharem com o(s) resultado(s) da mineração de dados, entreviste-os: Eles estão satisfeitos? O que poderia ter sido feito melhor? Eles precisam de suporte adicional?
ÿ Resumir o feedback e escrever a documentação da experiência ÿ Analisar o processo (coisas que funcionaram bem, erros cometidos, lições aprendidas, etc.) ÿ Documentar o processo de mineração de dados específico (como os resultados e a experiência de aplicar o modelo ser realimentado no processo?) ÿ
Generalizar a partir dos detalhes para tornar a experiência útil para projetos futuros

IV As saídas do CRISP-DM

Esta seção contém breves descrições da finalidade e do conteúdo dos relatórios mais importantes. Aqui, nos concentramos em relatórios destinados a comunicar os resultados de uma fase para pessoas não envolvidas nessa fase (e possivelmente não envolvidas neste projeto). Elas não são necessariamente idênticas às saídas descritas no modelo de referência e no guia do usuário. O objetivo dessas saídas é principalmente documentar os resultados durante a execução do projeto.

1 Compreensão do negócio

Os resultados da fase de Entendimento do negócio podem ser resumidos em um relatório. Sugerimos as seguintes seções:

Background A
seção Background fornece uma visão geral básica do contexto do projeto. Isso lista em que área o projeto está trabalhando, quais problemas foram identificados e por que a mineração de dados parece fornecer uma solução.

Objetivos de negócios e critérios de sucesso A
seção Objetivos de negócios descreve as metas do projeto em termos de negócios. Para cada objetivo, devem ser fornecidos Critérios de Sucesso Empresarial, ou seja, medidas explícitas para determinar se o projeto teve sucesso ou não em seus objetivos.
Esta seção também deve listar os objetivos que foram considerados, mas rejeitados. A justificativa da seleção dos objetivos deve ser dada.

Inventário de recursos A
seção Inventário de Recursos visa identificar pessoal, fontes de dados, instalações técnicas e outros recursos que possam ser úteis na execução do projeto.

Requisitos, suposições e restrições Esta seção lista
os requisitos gerais para a execução do projeto: tipo de resultados do projeto, suposições feitas sobre a natureza do problema e os dados que estão sendo usados e restrições impostas ao projeto.

Riscos e contingências
Esta seção identifica problemas que podem ocorrer no projeto, descreve as consequências e indica quais ações podem ser tomadas para minimizar tais riscos.

Terminologia
A seção Terminologia permite que pessoas não familiarizadas com os problemas abordados pelo projeto se familiarizem com eles.

Custos e benefícios

Esta seção descreve os custos do projeto e os benefícios comerciais previstos se o projeto for bem-sucedido (por exemplo, retorno sobre o investimento). Outros benefícios menos tangíveis (por exemplo, satisfação do cliente) também devem ser destacados.

Objetivos de mineração de dados e critérios de

sucesso A seção Objetivos de mineração de dados apresenta os resultados do projeto que permitem atingir os objetivos de negócios. Além de listar as prováveis abordagens de mineração de dados, também devem ser listados os critérios de sucesso para os resultados em termos de mineração de dados.

Plano do

projeto Esta seção lista as etapas a serem executadas no projeto, juntamente com sua duração, recursos necessários, entradas, saídas e dependências. Sempre que possível, deve explicitar as iterações em larga escala no processo de mineração de dados – por exemplo, repetições das fases de modelagem e avaliação.

Avaliação inicial de ferramentas e técnicas Esta

seção fornece uma visão inicial de quais ferramentas e técnicas provavelmente serão usadas e como. Ele descreve os requisitos para ferramentas e técnicas, lista as ferramentas e técnicas disponíveis e as compara com os requisitos.

2 Compreensão dos

dados Os resultados da fase de Compreensão dos dados geralmente são documentados em vários relatórios. Idealmente, esses relatórios devem ser escritos durante a execução das respectivas tarefas. Os relatórios descrevem os conjuntos de dados que são explorados durante a compreensão dos dados. Para o relatório final, basta um resumo das partes mais relevantes.

Relatório inicial de coleta de

dados Este relatório descreve como as diferentes fontes de dados identificadas no inventário foram capturadas e extraídas.

Tópicos a serem

abordados: ÿ Histórico dos

dados ÿ Lista de fontes de dados com ampla área de dados necessários cobertos por

cada ÿ Para cada fonte de dados, método de aquisição ou extração ÿ

Problemas encontrados na aquisição ou extração de dados

Relatório de descrição de dados

Cada conjunto de dados adquirido é descrito neste relatório.

Tópicos a serem abordados:

ÿ Cada fonte de dados descrita em detalhes

ÿ Lista de tabelas (pode ser apenas uma) ou outros objetos de banco de dados ÿ

Descrição de cada campo, incluindo unidades, códigos usados, etc.

Relatório de exploração de

dados Este relatório descreve a exploração de dados e seus resultados.

Tópicos a serem abordados:

ÿ Histórico, incluindo os objetivos gerais da exploração de dados. Para cada área de exploração realizada:

- – Regularidades ou padrões esperados
- – Método de detecção
- – Regularidades ou padrões encontrados, esperados e inesperados
- – Quaisquer outras surpresas
- – Conclusões para transformação de dados, limpeza de dados e qualquer outro pré-processamento
- – Conclusões relacionadas a metas de mineração de dados ou objetivos de negócios
- – Resumo das conclusões

Relatório de qualidade de

dados Este relatório descreve a abrangência e precisão dos dados.

Tópicos a serem abordados:

ÿ Histórico, incluindo amplas expectativas sobre a qualidade dos dados. Para cada conjunto de dados:

- – Abordagem adotada para avaliar a qualidade dos dados
- – Resultados da avaliação da qualidade dos dados
- – Resumo das conclusões de qualidade de dados

3 Preparação de dados

Os relatórios na fase de preparação de dados concentram-se nas etapas de pré-processamento que produzem os dados a serem minerados.

Relatório de descrição do

conjunto de dados Este relatório fornece uma descrição do conjunto de dados (após o pré-processamento) e o processo pelo qual foi produzido.

Tópicos a serem abordados:

ÿ Histórico, incluindo metas amplas e plano para pré-processamento ÿ Justificativa

para inclusão/exclusão de conjuntos de dados. Para cada conjunto de dados incluído:

– Descrição do pré-processamento, incluindo as ações necessárias para resolver quaisquer problemas de qualidade de dados

– Descrição detalhada do conjunto de dados resultante, tabela por tabela e campo por campo

– Justificativa para inclusão/exclusão de atributos

– Descobertas feitas durante o pré-processamento e quaisquer implicações para trabalhos futuros

- Sumário e conclusões

4 Modelagem

As saídas produzidas durante a fase de Modelagem podem ser combinadas em um relatório. Sugerimos as seguintes seções:

Suposições de modelagem

Esta seção define quaisquer suposições explícitas feitas sobre os dados e quaisquer suposições implícitas na técnica de modelagem a ser usada.

Projeto de

teste Esta seção descreve como os modelos são construídos, testados e avaliados.

Tópicos a serem cobertos:

ÿ Histórico - descreve a modelagem realizada e sua relação com os objetivos de mineração de dados. Para cada tarefa de modelagem:

– Descrição ampla do tipo de modelo e dos dados de treinamento a serem usados

– Explicação de como o modelo será testado ou avaliado

– Descrição de quaisquer dados necessários para o teste

– Planeje a produção de dados de teste, se houver

– Descrição de qualquer exame planejado de modelos por especialistas de domínio ou dados

– Resumo do plano de teste

Descrição do modelo

Este relatório descreve os modelos entregues e uma visão geral do processo pelo qual eles foram produzidos.

Tópicos a serem
abordados: ÿ Visão geral dos modelos produzidos. Para cada modelo:
– Tipo de modelo e relacionamento com os objetivos de mineração
de dados – Configurações de parâmetros usadas para
produzir o modelo – Descrição detalhada do modelo e quaisquer recursos especiais. Por exemplo:
ÿ Para modelos baseados em regras, liste as regras produzidas mais qualquer avaliação de precisão e cobertura por regra ou modelo geral ÿ Para
modelos opacos, liste todas as informações técnicas sobre o modelo (como topologia de rede neural) e qualquer
descrições comportamentais produzidas pelo processo de modelagem (como precisão ou sensibilidade)
ÿ Descrição do comportamento e interpretação do modelo
– Conclusões sobre padrões nos dados (se houver). Às vezes, o modelo revelará fatos importantes sobre os dados
sem um processo de avaliação separado (por exemplo, que a saída ou conclusão é duplicada em uma das entradas).
ÿ Resumo das conclusões

Avaliação do modelo

Esta seção descreve os resultados dos testes dos modelos de acordo com o projeto de teste.

Tópicos a serem
abordados: ÿ Visão geral do processo de avaliação e resultados, incluindo quaisquer desvios do plano de teste. Para cada modelo: –
Avaliação detalhada, incluindo medições como precisão e interpretação do comportamento – Quaisquer comentários
sobre modelos por especialistas em domínio ou dados – Avaliação
resumida de modelos – Insights sobre por
que uma determinada técnica de modelagem e certas configurações de parâmetros levaram a bons/maus resultados - Avaliação
resumida do conjunto completo de modelos

5 Avaliação

Avaliação dos resultados da mineração de dados com relação aos critérios de sucesso do negócio Este relatório compara os resultados da mineração de dados com os objetivos de negócios e os critérios de sucesso do negócio.

Tópicos a serem abordados:

ÿ Revisão dos objetivos de negócios e critérios de sucesso de negócios (que podem ter mudado durante e/ou como resultado de mineração de dados). Para cada critério de sucesso do negócio:

– Comparação detalhada entre o critério de sucesso e os resultados da mineração de dados –

Conclusões sobre a viabilidade do critério de sucesso e adequação do processo de mineração de dados

ÿ Revisão do sucesso do projeto:

– O projeto alcançou os objetivos de negócios originais?

– Existem novos objetivos de negócios a serem abordados posteriormente no projeto ou em novos projetos?

– Conclusões para futuros projetos de mineração de dados

Revisão do processo

Esta seção avalia a eficácia do projeto e identifica quaisquer fatores que podem ter sido negligenciados e que devem ser levados em consideração se o projeto for repetido.

Lista de possíveis ações

Esta seção faz recomendações sobre as próximas etapas do projeto.

6 Implantação

Plano de implantação

Este relatório especifica a implantação dos resultados da mineração de dados.

Tópicos a serem abordados:

ÿ Resumo dos resultados implantáveis (derivados do relatório Próximas etapas) ÿ

Descrição do plano de implantação

Plano de monitoramento e manutenção O

plano de monitoramento e manutenção especifica como os resultados implantados devem ser mantidos. Tópicos a serem abordados: ÿ Visão geral da implantação dos resultados e indicação de quais resultados podem exigir atualização (e por quê). Para cada resultado implantado:

– Descrição de como a atualização será acionada (atualizações regulares, evento de acionamento, monitoramento de desempenho)

– Descrição de como será realizada a atualização

ÿ Resumo do processo de atualização de resultados

Relatório final

O relatório final é usado para resumir o projeto e seus resultados.

Conteúdo:

ÿ Resumo do entendimento do negócio: plano de fundo, objetivos e critérios de sucesso ÿ Resumo do processo de mineração de dados ÿ Resumo dos resultados da mineração de dados ÿ Resumo da avaliação dos resultados ÿ Resumo dos planos de implantação e manutenção ÿ Análise de custo/benefício
ÿ Conclusões para o negócio

ÿ Conclusões para mineração de dados futura

7 Resumo das dependências

A tabela a seguir resume as principais entradas para as entregas. Isso não significa que apenas as entradas listadas devem ser consideradas - por exemplo, os objetivos de negócios devem ser difundidos em todas as entregas. No entanto, as entregas devem abordar questões específicas levantadas por suas contribuições.

Estágio	Entregável	Refere-se a	Intimamente relacionado com
Negócios	Fundo		
Entendimento	Objetivos de negócios	Fundo	Terminologia
	Critérios de sucesso comercial	Objetivos de negócios	
	Inventário de Recursos		
	Requisitos, Suposições e Restrições	Objetivos de negócios	
	Riscos & contingências	Objetivos de Negócios; Critérios de sucesso comercial	
	Terminologia	Fundo	Objetivos de negócios
	Custos e benefícios	Objetivos de negócios	Plano de projeto
	Metas de Mineração de Dados	Objetivos de Negócios; Requisitos, Suposições e Restrições	
	Critérios de sucesso da mineração de dados	Critérios de Sucesso Empresarial; Requisitos; Premissas & Restrições; Metas de Mineração de Dados	
	Plano de projeto	Objetivos de Negócios; Inventário de Recursos; Requisitos; Premissas & Restrições; Riscos e Contingências	Custos e Benefícios
Entendimento de dados	Relatório inicial de coleta de dados	Metas de Negócios; Inventário de Recursos; Metas de Mineração de Dados	
	Relatório de descrição de dados	Metas de Negócios; Relatório inicial de coleta de dados	Relatório de qualidade de dados
	Relatório de qualidade de dados	Metas de Negócios; Relatório inicial de coleta de dados	Relatório de descrição de dados
	Relatório de Análise Exploratória	Metas de Negócios; Relatório inicial de coleta de dados	
Preparação de dados	Conjunto de dados e descrição do conjunto de dados	Metas de Negócios; Metas de mineração de dados e Relatório de descrição de dados; Qualidade de Dados Relatório; Relatório de Análise Exploratória	
Modelagem	Projeto de teste	Metas de Mineração de Dados; Critérios de sucesso da mineração de dados	
	modelos	Metas de Mineração de Dados	Configurações de Parâmetros
	Configurações de Parâmetros	Modelos de metas de	modelos
	Descrição do modelo	mineração de dados; Configurações de Parâmetros;	
	Avaliação	Critérios de sucesso de mineração de dados de projeto de teste;	
Avaliação	Nota de avaliação Critérios de sucesso comercial	Projeto de teste; Modelos de Critérios de Sucesso	
	Revisão do Processo	Empresarial; Terminologia Metas de Negócios; Nota de avaliação Critérios de sucesso comercial	
	Próximos passos	Plano de projeto; Nota de avaliação Critérios de sucesso comercial	
Implantação	Plano de preparação	Metas de Negócios; Requisitos, Suposições e Restrições Metas de Negócios;	Plano de manutenção
	Plano de manutenção	Requisitos, Suposições e Restrições Metas de Negócios; Terminologia; Avaliação	Plano de preparação
	Relatório Final e Apresentação	wrt Critérios de sucesso de negócios	
	Documentação da experiência	Plano de projeto; Revisão do Processo	

V Apêndice

1 Glossário/terminologia

Atividade – Parte de uma tarefa no Guia do Usuário; descreve ações para realizar uma tarefa

Metodologia CRISP-DM - O termo geral para todos os conceitos desenvolvidos e definidos no CRISP-DM

Contexto de mineração de dados – Um conjunto de restrições e suposições, como tipo de problema, técnicas ou ferramentas, domínio de aplicativo

Tipo de problema de mineração de dados – Uma classe de problemas típicos de mineração de dados, como descrição e resumo de dados, segmentação, descrições de conceito, classificação, previsão, análise de dependência

Genérico – Uma tarefa que se mantém em todos os projetos de mineração de dados possíveis

Modelo – A capacidade de aplicar algoritmos a um conjunto de dados para prever atributos de destino; executável

Saída – O resultado tangível da execução de uma tarefa

Fase – Um termo para a parte de alto nível do modelo de processo CRISP-DM; consiste em tarefas relacionadas

Instância de processo - Um projeto específico descrito em termos do modelo de processo

Modelo de processo – Define a estrutura de projetos de mineração de dados e fornece orientação para sua execução; consiste em modelo de referência e guia do usuário

Modelo de referência – Decomposição de projetos de mineração de dados em fases, tarefas e saídas

Especializado - Uma tarefa que faz suposições específicas em contextos específicos de mineração de dados

Tarefa – Uma série de atividades para produzir um ou mais resultados; parte de uma fase

Guia do usuário – Conselhos específicos sobre como realizar projetos de mineração de dados

2 Tipos de problemas de mineração de dados

Normalmente, o projeto de mineração de dados envolve uma combinação de diferentes tipos de problemas, que juntos resolvem o problema de negócios.

2.1 Descrição e sumarização dos dados

A descrição e sumarização dos dados visa a descrição concisa das características dos dados, normalmente de forma elementar e agregada. Isso dá ao usuário uma visão geral da estrutura dos dados. Às vezes, apenas a descrição e o resumo dos dados podem ser um objetivo de um projeto de mineração de dados. Por exemplo, um varejista pode estar interessado no volume de negócios de todos os pontos de venda discriminados por categorias. Mudanças e diferenças em um período anterior podem ser resumidas e destacadas. Esse tipo de problema estaria na extremidade inferior da escala dos problemas de mineração de dados.

Em quase todos os projetos de mineração de dados, no entanto, a descrição e o resumo dos dados são uma meta secundária no processo, normalmente em seus estágios iniciais. No início de um processo de mineração de dados, o usuário geralmente não conhece nem o objetivo preciso da análise nem a natureza precisa dos dados. A análise exploratória inicial de dados pode ajudar os usuários a entender a natureza dos dados e a formar hipóteses potenciais para informações ocultas. Técnicas simples de estatísticas descritivas e de visualização fornecem os primeiros insights sobre os dados. Por exemplo, a distribuição de clientes por idade e regiões geográficas sugere quais partes de um grupo de clientes precisam ser abordadas por outras estratégias de marketing.

A descrição e o resumo dos dados geralmente ocorrem em combinação com outros tipos de problemas de mineração de dados. Por exemplo, a descrição de dados pode levar à postulação de segmentos interessantes nos dados. Depois que os segmentos são identificados e definidos, uma descrição e um resumo desses segmentos são úteis. É aconselhável realizar a descrição e o resumo dos dados antes de abordar qualquer outro tipo de problema de mineração de dados. Neste documento, isso se reflete no fato de que a descrição e resumo dos dados é uma tarefa na fase de compreensão dos dados.

A sumarização também desempenha um papel importante na apresentação dos resultados finais. Os resultados dos outros tipos de problemas de mineração de dados (por exemplo, descrições de conceitos ou modelos de previsão) também podem ser considerados resumos de dados, mas em um nível conceitual superior.

Muitos sistemas de relatório, pacotes estatísticos, OLAP e sistemas EIS podem abranger a descrição e o resumo de dados, mas geralmente não fornecem nenhum método para realizar modelagem mais avançada. Se a descrição e o resumo dos dados forem considerados um tipo de problema autônomo e nenhuma modelagem adicional for necessária, essas ferramentas podem ser apropriadas para realizar a mineração de dados compromissos.

2.2 Segmentação

A segmentação visa a separação dos dados em subgrupos ou classes interessantes e significativos. Todos os membros de um subgrupo compartilham características comuns. Por exemplo, na análise de cestas de compras, pode-se definir segmentos de cestas dependendo dos itens que elas contêm.

A segmentação pode ser realizada de forma manual ou semi-automática. O analista pode hipotetizar certos subgrupos como relevantes para a questão do negócio, com base no conhecimento prévio ou no resultado da descrição e resumo dos dados. Além disso, também existem técnicas de agrupamento automático que podem detectar estruturas anteriormente insuspeitas e ocultas nos dados que permitem a segmentação.

Às vezes, a segmentação pode ser um objetivo de mineração de dados. Então a detecção de segmentos seria o objetivo principal de um projeto de mineração de dados. Por exemplo, todos os endereços em áreas de CEP com idade e renda acima da média podem ser selecionados para envio de anúncios de seguro residencial para idosos.

Muitas vezes, no entanto, a segmentação é um passo para resolver outros tipos de problemas. Então, o objetivo é manter o tamanho dos dados gerenciáveis ou encontrar subconjuntos de dados homogêneos que sejam mais fáceis de analisar. Normalmente, em grandes conjuntos de dados, várias influências se sobrepõem e obscurecem os padrões interessantes. Então, a segmentação apropriada torna a tarefa mais fácil. Por exemplo, analisar dependências entre itens em milhões de cestas de compras é muito difícil. É muito mais fácil (e mais significativo, normalmente) identificar dependências em segmentos interessantes de cestas de compras - por exemplo, cestas de alto valor, cestas contendo produtos de conveniência ou cestas de um determinado dia ou hora.

Nota: Na literatura, existe alguma ambiguidade no significado de certos termos. Às vezes, a segmentação é chamada de agrupamento ou classificação. O último termo é confuso porque algumas pessoas o usam para se referir à criação de classes, enquanto outros significam a criação de modelos para prever classes conhecidas para casos não vistos anteriormente. Neste documento, restringimos o termo classificação ao último significado (veja abaixo) e usamos o termo segmentação para o primeiro significado, embora técnicas de classificação possam ser usadas para obter descrições dos segmentos descobertos.

Técnicas apropriadas: ÿ
Técnicas de agrupamento
ÿ Redes neurais
ÿ Visualização

Exemplo:

Uma empresa automobilística coleta regularmente informações sobre seus clientes sobre suas características socioeconômicas, como renda, idade, sexo, profissão, etc. Usando a análise de cluster, a empresa pode dividir seus clientes em subgrupos mais compreensíveis e analisar a estrutura de cada subgrupo. Estratégias de marketing específicas são implantadas para cada grupo separado.

2.3 Descrições de conceitos

A descrição de conceitos visa uma descrição compreensível de conceitos ou classes. O objetivo não é desenvolver modelos completos com alta precisão de previsão, mas obter insights. Por exemplo, uma empresa pode estar interessada em saber mais sobre seus clientes leais e desleais. A partir de uma descrição conceitual desses conceitos (clientes leais e desleais) a empresa pode inferir o que poderia ser feito para manter os clientes leais ou para transformar clientes desleais em clientes leais.

A descrição do conceito tem uma conexão estreita com a segmentação e a classificação. A segmentação pode levar a uma enumeração de objetos pertencentes a um conceito ou classe sem fornecer nenhuma descrição compreensível. Normalmente, a segmentação é realizada antes da descrição do conceito. Algumas técnicas - técnicas de agrupamento conceitual, por exemplo - executam a segmentação e a descrição do conceito ao mesmo tempo.

Descrições de conceito também podem ser usadas para fins de classificação. Por outro lado, algumas técnicas de classificação produzem modelos de classificação compreensíveis, que podem ser considerados descrições de conceitos. A distinção importante é que a classificação visa ser completa em algum sentido. O modelo de classificação precisa ser aplicado a todos os casos na população selecionada.
Por outro lado, as descrições de conceitos não precisam ser completas. É suficiente que descrevam partes importantes dos conceitos ou classes. No exemplo acima, pode ser suficiente obter descrições conceituais dos clientes que são claramente leais.

Técnicas apropriadas:

ÿ Métodos de indução de regras

ÿ Agrupamento conceitual

Exemplo:

Usando dados sobre os compradores de carros novos e uma técnica de indução de regras, uma montadora pode gerar regras que descrevam seus clientes leais e desleais. Seguem exemplos das regras geradas: *SEX = masculino e*

Se IDADE > 51 então CLIENTE = leal SEXO = feminino e IDADE

Se > 21 então CLIENTE = leal PROFISSÃO = gerente e IDADE <

Se 51 então CLIENTE = desleal STATUS FAMILIAR = solteiro e IDADE < 51 então

Se CLIENTE = desleal

2.4 Classificação A

classificação pressupõe que existe um conjunto de objetos caracterizados por alguns atributos ou feições pertencentes a diferentes classes. O rótulo da classe é um valor discreto (simbólico) e é conhecido para cada objeto. O objetivo é construir modelos de classificação (às vezes chamados de classificadores), que atribuem o rótulo de classe correto a objetos não vistos e não rotulados anteriormente.
Os modelos de classificação são usados principalmente para modelagem preditiva.

Os rótulos de classe podem ser fornecidos antecipadamente — definidos pelo usuário, por exemplo, ou derivados da segmentação. A classificação é um dos tipos de problemas de mineração de dados mais importantes que ocorre em uma ampla gama de aplicações. Muitos problemas de mineração de dados podem ser transformados em problemas de classificação. Por exemplo, a pontuação de crédito tenta avaliar o risco de crédito de um novo cliente.
Isso pode ser transformado em um problema de classificação criando duas classes, bons e maus clientes. Um modelo de classificação pode ser gerado a partir de dados existentes do cliente sobre seu comportamento de crédito. Esse modelo de classificação pode então ser usado para atribuir novos clientes a uma das duas classes e aceitá-los ou rejeitá-los.

A classificação tem conexões com quase todos os outros tipos de problemas. Problemas de predição podem ser transformados em problemas de classificação por binning de rótulos de classes contínuas, uma vez que as técnicas de binning permitem transformar faixas contínuas em intervalos discretos. Esses intervalos discretos, em vez dos valores numéricos exatos, são usados como rótulos de classe e, portanto, levam a um problema de classificação. Algumas técnicas de classificação produzem descrições compreensíveis de classes ou conceitos. Há também uma conexão com a análise de dependência porque os modelos de classificação normalmente exploram e elucidam as dependências entre
atributos.

A segmentação pode fornecer os rótulos de classe ou restringir o conjunto de dados para que bons modelos de classificação possam ser construídos. É útil analisar os desvios antes de construir um modelo de classificação. Desvios e outliers podem obscurecer os padrões que permitiriam um bom modelo de classificação. Por outro lado, um modelo de classificação também pode ser usado para identificar desvios e outros problemas com os dados.

Técnicas apropriadas: ÿ Análise
discriminante
ÿ Métodos de indução de regras
ÿ Aprendizagem da árvore de decisão
ÿ Redes neurais
ÿ K vizinho mais próximo ÿ
Raciocínio baseado em casos ÿ
Algoritmos genéticos

Exemplo:

Os bancos geralmente têm informações sobre o comportamento de pagamento de seus solicitantes de crédito. Combinando essas informações financeiras com outras informações sobre os clientes, como sexo, idade, renda, etc., é possível desenvolver um sistema para classificar novos clientes como bons ou maus clientes (ou seja, o risco de crédito na aceitação de um cliente é Alto ou baixo).

2.5 Previsão

Outro tipo de problema importante que ocorre em uma ampla gama de aplicações é a previsão. A previsão é muito semelhante à classificação.
A única diferença é que na predição o atributo alvo (classe) não é um atributo qualitativo discreto, mas contínuo.
O objetivo da previsão é encontrar o valor numérico do atributo de destino para objetos invisíveis. Na literatura, esse tipo de problema às vezes é chamado de regressão. Se a previsão lida com dados de séries temporais, geralmente é chamada de previsão.

Técnicas apropriadas: ÿ
Análise de regressão
ÿ Árvores de regressão
ÿ Redes neurais
ÿ K vizinho mais próximo
ÿ Métodos Box-Jenkins ÿ
Algoritmos genéticos

Exemplo:

A receita anual de uma empresa internacional está correlacionada com outros atributos como publicidade, taxa de câmbio, taxa de inflação, etc. Com esses valores (ou estimativas confiáveis), a empresa pode prever sua receita esperada para o próximo ano.

2.6 Análise de dependência A

análise de dependência consiste em encontrar um modelo que descreva dependências significativas (ou associações) entre itens de dados ou eventos. As dependências podem ser usadas para prever o valor de um item de dados com informações sobre outros itens de dados. Embora as dependências possam ser usadas para modelagem preditiva, elas são usadas principalmente para compreensão. As dependências podem ser estritas ou probabilísticas.

As associações são um caso especial de dependências, que recentemente se tornaram muito populares. As associações descrevem afinidades de itens de dados (isto é, itens de dados ou eventos que frequentemente ocorrem juntos). Um cenário típico de aplicação para associações é a análise de cestas de compras. Lá, uma regra como “em 30% de todas as compras, cerveja e amendoim foram comprados juntos” é um exemplo típico de associação.

Algoritmos para detectar associações são muito rápidos e produzem muitas associações. Selecionar os mais interessantes é um desafio.

A análise de dependência tem conexões estreitas com previsão e classificação, pois as dependências são usadas implicitamente para a formulação de modelos preditivos. Há também uma conexão com as descrições de conceito, que geralmente destacam as dependências.

Em aplicativos, a análise de dependência geralmente ocorre simultaneamente com a segmentação. Em grandes conjuntos de dados, as dependências raramente são significativas porque muitas influências se sobrepõem. Nesses casos, é aconselhável realizar uma análise de dependência em segmentos mais homogêneos dos dados.

Padrões sequenciais são um tipo especial de dependências nas quais a ordem dos eventos é considerada. Em uma análise de cesta de compras, as associações descrevem as dependências entre os itens em um determinado momento. Os padrões sequenciais descrevem os padrões de compras de um cliente específico ou de um grupo de clientes ao longo do tempo.

Técnicas apropriadas: ÿ
Análise de correlação ÿ
Análise de regressão
ÿ Regras de associação
ÿ Redes bayesianas ÿ
Programação lógica indutiva ÿ
Técnicas de visualização

Exemplo 1:

Usando a análise de regressão, um analista de negócios descobriu que existem dependências significativas entre as vendas totais de um produto e seu preço e o valor gasto em publicidade. Esse conhecimento permite ao negócio atingir o nível desejado de vendas, alterando o preço do produto e/ou o gasto com propaganda.

Exemplo 2:

Aplicando algoritmos de regras de associação a dados sobre acessórios de automóveis, uma montadora descobriu que, em 95% dos casos, se um CD player for solicitado, uma transmissão automática também será solicitada. A partir dessa dependência, a montadora decide oferecer esses acessórios em um pacote, o que leva à redução de custos.

SPSS

Para saber mais, visite www.spss.com. Para obter os endereços dos escritórios e números de telefone da SPSS, acesse www.spss.com/worldwide.