Guida CRISP-DM di IBM SPSS Modeler

Nota

Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni presenti in “Note” a pagina 39.

Informazioni sul prodotto

Questa edizione si applica alla versione 18, release 2, livello di modifica 2 di IBM® SPSS Modeler e a tutte le successive release e modifiche a meno che non sia diversamente indicato nelle nuove edizioni.

© Copyright International Business Machines Corporation .

Indice

Prefazione ..... **vii**

Capitolo 1. Introduzione a CRISP-DM **1**
Panoramica su CRISP-DM 1
CRISP-DM in IBM SPSS Modeler 2
Risorse aggiuntive 3

Capitolo 2. Business Understanding **5**
Panoramica sulla fase Business Understanding 5
Individuazione degli obiettivi di business 5
Esempio di vendita al dettaglio in linea: individuazione degli obiettivi di business 5
Definizione della situazione di business 5
Definizione degli obiettivi di business 6
Criteri per la valutazione dell'esito del business 6
Valutazione della situazione 7
Esempio di vendita al dettaglio in linea: valutazione della situazione 7
Inventario delle risorse 7
Requisiti, presupposti e vincoli 8
Rischi e imprevisti 8
Terminologia 9
Analisi dei costi/benefici 9
Definizione degli obiettivi di data mining 9
Obiettivi del data mining 9
Esempio di vendita al dettaglio in linea: obiettivi di data mining 10
Criteri per la valutazione dell'esito positivo del data mining 10
Produzione di un piano di progetto 10
Stesura del piano di progetto 10
Esempio di piano di progetto 11
Valutazione di strumenti e tecniche 11
Considerazioni utili 11

Capitolo 3. Data Understanding **13**
Panoramica sulla fase Data Understanding 13
Raccolta di dati iniziali 13
Esempio di vendita al dettaglio in linea: raccolta di dati iniziali 13
Creazione di un report di raccolta dei dati 14
Descrizione dei dati 14
Esempio di vendita al dettaglio in linea: descrizione dei dati 14
Creazione di un report di descrizione dei dati 15
Analisi dei dati 15
Esempio di vendita al dettaglio in linea: analisi dei dati 15
Creazione di un report di esplorazione dati 16
Verifica della qualità dei dati 16
Esempio di vendita al dettaglio in linea: verifica della qualità dei dati 16
Creazione di un report sulla qualità dei dati 17
Considerazioni utili 17

Capitolo 4. Data Preparation **19**
Panoramica sulla fase Data Preparation 19
Selezione di dati 19

Esempio di vendita al dettaglio in linea: selezione dei dati..... 19
Inclusione o esclusione dei dati..... 19
Clean Data..... 20
Esempio di vendita al dettaglio in linea: pulitura dei dati..... 20
Creazione di un report di pulitura dei dati..... 21
Creazione di nuovi dati..... 21
Esempio di vendita al dettaglio in linea: creazione di dati..... 21
Derivazione di attributi..... 21
Integrazione dei dati..... 22
Esempio di vendita al dettaglio in linea: integrazione dei dati..... 22
Attività di integrazione..... 22
Formattazione dei dati..... 23
Considerazioni per la modellazione..... 23

Capitolo 5. Modellazione..... 25
Panoramica sulla fase di modellazione..... 25
Selezione delle tecniche di modellazione..... 25
Esempio di vendita al dettaglio in linea: tecniche di modellazione..... 25
Scelta delle tecniche di modellazione appropriate..... 26
Presupposti della modellazione..... 26
Generazione di un progetto di test..... 26
Creazione di un progetto di test..... 26
Esempio di vendita al dettaglio in linea: progetto di test..... 27
Creazione dei modelli..... 27
Esempio di vendita al dettaglio in linea: creazione del modello..... 27
Impostazioni dei parametri..... 28
Esecuzione dei modelli..... 28
Descrizione del modello..... 28
Valutazione del modello..... 28
Valutazione completa del modello..... 28
Esempio di vendita al dettaglio in linea: valutazione dei modelli..... 29
Registrazione dei parametri modificati..... 29
Considerazioni utili..... 29

Capitolo 6. Valutazione..... 31
Panoramica sulla fase di valutazione..... 31
Valutazione dei risultati..... 31
Esempio di vendita al dettaglio in linea: valutazione dei risultati..... 31
Rivedi processo..... 32
Esempio di vendita al dettaglio in linea: report di verifica..... 32
Individuazione dei passaggi successivi..... 32
Esempio di vendita al dettaglio in linea: passaggi successivi..... 33

Capitolo 7. Distribuzione..... 35
Panoramica sulla fase Distribuzione..... 35
Pianificazione della distribuzione..... 35
Esempio di vendita al dettaglio in linea: pianificazione della distribuzione..... 35
Pianificazione delle attività di monitoraggio e manutenzione..... 36
Esempio di vendita al dettaglio in linea: monitoraggio e manutenzione..... 36
Creazione di un report finale..... 37
Preparazione di una presentazione finale..... 37
Esempio di vendita al dettaglio in linea: report finale..... 37
Verifica finale del progetto..... 38
Esempio di vendita al dettaglio in linea: verifica finale..... 38

Note..... 39
Marchi..... 40

Termini e condizioni per la documentazione del prodotto..... 40

Indice analitico..... 43

Prefazione

IBM SPSS Modeler è l'efficace workbench di data mining aziendale di IBM Corp.. SPSS Modeler consente alle organizzazioni di migliorare le relazioni con i clienti e con il pubblico grazie a un'analisi approfondita dei dati. Le organizzazioni potranno utilizzare le informazioni ottenute tramite SPSS Modeler per mantenere i clienti di valore, cogliere opportunità di vendite incrociate, attrarre nuovi clienti, individuare frodi, diminuire i rischi e migliorare l'offerta di servizi a livello statale.

L'interfaccia visiva di SPSS Modeler invita gli utenti ad applicare la propria competenza di business specifica, con cui sarà possibile ottenere modelli di previsione più efficaci e una riduzione nei tempi di sviluppo delle soluzioni. SPSS Modeler offre una vasta gamma di tecniche di creazione di modelli, quali previsione, classificazione, segmentazione e algoritmi per l'individuazione delle associazioni. IBM SPSS Modeler Solution Publisher consente quindi di distribuire a livello aziendale i modelli creati in modo che vengano utilizzati dai responsabili dei processi decisionali oppure inseriti in un database.

Informazioni su IBM Business Analytics

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni di business. Un ampio portafoglio di applicazioni di business intelligence, analisi predittiva, gestione delle prestazioni e delle strategie finanziarie e analisi offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività di business. Le aziende, gli enti governativi e le università di tutto il modo si affidano alla tecnologia IBM SPSS perché rappresenta un vantaggio concorrenziale in termini di attrazione, retention ed aumento dei clienti, riducendo al tempo stesso le frodi e limitando i rischi. Incorporando il software IBM SPSS nelle attività quotidiane, le aziende diventano imprese in grado di effettuare previsioni e di gestire ed automatizzare le decisioni, per raggiungere gli obiettivi di business e vantaggi tangibili sulla concorrenza. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito http://www.ibm.com/spss.

Supporto tecnico

Il supporto tecnico è a disposizione dei clienti che dispongono di un contratto di manutenzione. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo di prodotti IBM o per l'installazione di uno degli ambienti hardware supportati. Per contattare il supporto tecnico, visitare il sito Web IBM all'indirizzo http://www.ibm.com/support. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del accordo di manutenzione.

Capitolo 1. Introduzione a CRISP-DM

Panoramica su CRISP-DM

CRISP-DM è l'acronimo di Cross-Industry Standard Process for Data Mining, un metodo di comprovata efficacia per l'esecuzione di operazioni di data mining.

- Come **metodologia**, comprende descrizioni delle tipiche fasi di un progetto e delle attività incluse in ogni fase e fornisce una spiegazione delle relazioni esistenti tra tali attività.
- Come **modello di elaborazione**, CRISP-DM fornisce una panoramica del ciclo di vita del data mining.

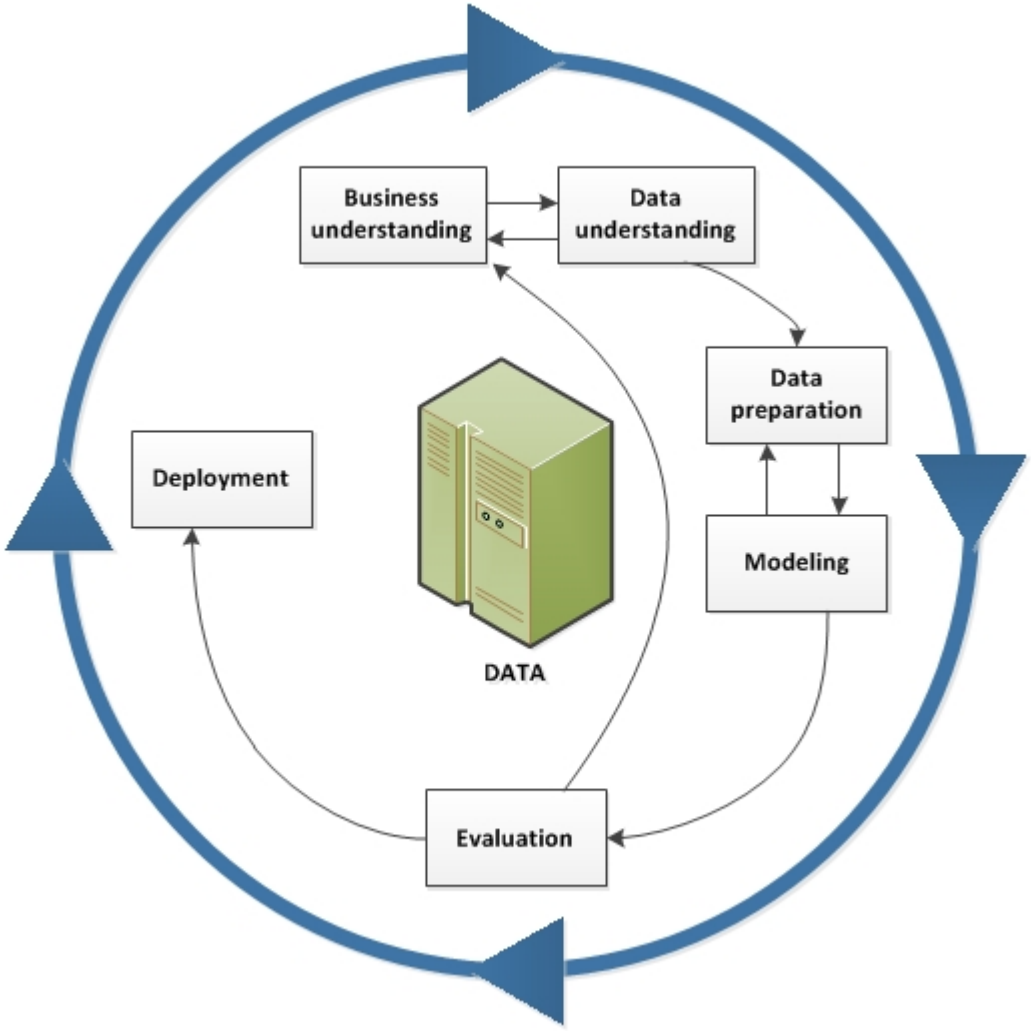

Figura 1. Ciclo di vita del data mining

Il modello del ciclo di vita è costituito da sei fasi con frecce che indicano le dipendenze più importanti e frequenti tra le diverse fasi. La sequenza delle fasi non è rigorosa. Nella maggior parte dei progetti, infatti, ci si sposta avanti e indietro tra le fasi in base alle necessità.

Il modello CRISP-DM è flessibile e può essere facilmente personalizzato. Se, ad esempio, l'azienda mira a rilevare il riciclaggio di denaro, è probabile che saranno passate in rassegna enormi quantità di dati senza uno specifico obiettivo di modellazione. Invece che sulla modellazione, il lavoro si concentrerà sulla visualizzazione e l'esplorazione dati al fine di individuare schemi sospetti nei dati finanziari. CRISP-DM consente di creare modelli di data mining adeguati alle diverse esigenze.

In una situazione come quella citata, le fasi di modellazione, valutazione e distribuzione potrebbero essere meno rilevanti di quelle di Data Understanding e Data Preparation. È tuttavia importante analizzare alcune delle domande emerse durante le tre ultime fasi del progetto per una pianificazione a lungo termine e futuri obiettivi di data mining.

CRISP-DM in IBM SPSS Modeler

In IBM SPSS Modeler la metodologia CRISP-DM è resa disponibile in due modi, per assicurare un supporto efficace al data mining.

- Lo strumento per i progetti di CRISP-DM consente di organizzare annotazioni, output e flussi di progetto in base alle fasi di un progetto tipico di data mining. Durante il progetto è possibile creare report in qualsiasi momento utilizzando le note relative a flussi e fasi CRISP-DM.
- La Guida in linea di CRISP-DM rappresenta una fonte di informazioni importante per la creazione di un progetto di data mining. Il sistema di guida include elenchi di attività per ogni passaggio, nonché esempi pratici del funzionamento di CRISP-DM. Per accedere alla Guida in linea di CRISP-DM, scegliere **Guida in linea di CRISP-DM** dal menu Aiuto della finestra principale.

Strumento per i progetti di CRISP-DM

Lo strumento per i progetti di CRISP-DM garantisce un approccio strutturato al data mining per assicurare la riuscita del progetto. Tale strumento rappresenta essenzialmente un'estensione dello strumento per i progetti standard di IBM SPSS Modeler. È infatti possibile passare dalla visualizzazione CRISP-DM alla visualizzazione Classi standard per verificare l'organizzazione di flussi e input in base alle fasi CRISP-DM o al tipo.

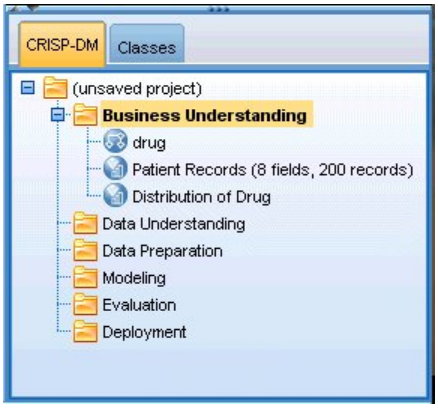

Figura 2. Strumento per i progetti di CRISP-DM

Tramite la visualizzazione CRISP-DM dello strumento per i progetti è possibile eseguire le operazioni seguenti:

- Organizzare un flusso di progetto e output in base alle fasi di data mining.
- Prendere note sugli obiettivi dell'organizzazione per ogni fase.
- Creare testi descrittivi personalizzati per ogni fase.
- Prendere note sulle conclusioni formulate in base a un determinato grafico o modello.
- Generare un report HTML o aggiornarlo per la distribuzione al team di progetto.

Guida in linea di CRISP-DM

IBM SPSS Modeler rende disponibile una Guida in linea per il modello di elaborazione CRISP-DM non proprietario. La Guida è organizzata in base alle fasi del progetto e include le informazioni di supporto seguenti:

- Panoramica ed elenco di attività per ogni fase CRISP-DM
- Informazioni sulla creazione di report per diversi tipi di operazioni cardine

- Esempi concreti che illustrano le modalità di utilizzo di CRISP-DM per semplificare i processi di data mining
- Collegamenti a risorse aggiuntive su CRISP-DM

Per accedere alla Guida in linea di CRISP-DM, scegliere **Guida in linea di CRISP-DM** dal menu Aiuto della finestra principale.

Risorse aggiuntive

Oltre a consultare la documentazione di supporto per CRISP-DM in IBM SPSS Modeler, è possibile approfondire la conoscenza dei processi di data mining in diversi modi.

- Consultare il manuale CRISP-DM, redatto dal consorzio CRISP-DM e fornito con la presente release.
- Consultare *Data Mining with Confidence*, copyright 2002 by SPSS Inc., ISBN 1-56827-287-1.

Capitolo 2. Business Understanding

Panoramica sulla fase Business Understanding

Prima di utilizzare IBM SPSS Modeler, è opportuno dedicare del tempo ad analizzare quali siano le aspettative dell'azienda nei confronti del data mining. Nell'indagine è opportuno coinvolgere quante più persone possibile e documentare i risultati. Il passaggio finale di questa fase di CRISP-DM descrive la procedura di elaborazione di un piano di progetto utilizzando le informazioni raccolte durante il processo di analisi.

Sebbene la ricerca possa sembrare superflua, non lo è affatto. Individuare quali sono le ragioni di business per le quali si decide di avvalersi del data mining è un buon modo per garantire che tutti i soggetti interessati procedano all'unisono prima di spendere risorse di valore.

Individuazione degli obiettivi di business

La prima attività consiste nel tentare di ottenere quante più informazioni è possibile sugli obiettivi di business da raggiungere mediante il data mining. Questa ricerca potrebbe non essere semplice come sembra, ma consente di ridurre al minimo eventuali rischi, facendo luce su problemi, obiettivi e risorse.

Il modello CRISP-DM garantisce una procedura strutturata per portare a termine l'attività descritta.

Elenco delle attività

- Iniziare a raccogliere informazioni di base sulla situazione di business corrente.
- Documentare gli obiettivi di business specifici definiti dai responsabili delle decisioni.
- Convenire sui criteri utilizzati per determinare l'esito del data mining dal punto di vista di business.

Esempio di vendita al dettaglio in linea: individuazione degli obiettivi di business

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Poiché sempre più aziende passano alla vendita sul Web, un affermato rivenditore in linea di computer e dispositivi elettronici si trova a dover affrontare la crescente concorrenza dei siti più recenti. Appurato che i punti vendita sul Web nascono con la stessa velocità, se non maggiore, con cui i clienti passano ad acquistare prodotti in linea, la società deve individuare un modo per restare redditizia nonostante l'aumento dei costi associati all'acquisizione di clienti. Una delle soluzioni proposte consiste nel coltivare le relazioni con i clienti esistenti, per massimizzare il valore di ognuno di loro.

Viene quindi commissionato uno studio con gli obiettivi seguenti:

- Aumentare le vendite incrociate migliorando le indicazioni.
- Fidelizzare ulteriormente i clienti grazie a un servizio più personalizzato.

In linea generale, l'esito dello studio sarà considerato positivo se:

- Le vendite incrociate aumenteranno del 10%.
- I clienti trascorreranno più tempo sul sito e visualizzeranno più pagine per visita.
- Lo studio si concluderà rispettando i limiti di tempo e di budget.

Definizione della situazione di business

Comprendere quale sia la situazione di business dell'azienda consente di conoscere lo scenario su cui si deve intervenire in termini di:

- Risorse disponibili (personale e materiali)

- Problemi
- Obiettivi

È necessario eseguire delle ricerche sulla situazione di business corrente, per individuare delle risposte reali alle domande che possono incidere sul risultato del progetto di data mining.

Attività 1: delineare la struttura aziendale

- Creare degli organigrammi per illustrare le divisioni, i reparti e i gruppi di progetti dell'azienda. includendo nomi e responsabilità dei manager.
- Identificare gli individui chiave dell'organizzazione.
- Identificare uno sponsor interno che fornisca supporto in termini finanziari e/o di esperienza specifica.
- Determinare se esiste un comitato direttivo e procurarsi l'elenco dei membri.
- Identificare le untià di business su cui inciderà il progetto di data mining.

Attività 2: descrivere l'area problematica

- Identificare l'area problematica, ad esempio il reparto marketing, assistenza clienti o sviluppo di business.
- Descrivere il problema in termini generali.
- Chiarire i prerequisiti del progetto, ad esempio le motivazioni per cui è stato avviato, se l'azienda si avvale già del data mining e così via.
- Verificare lo stato del progetto di data mining all'interno del gruppo di business, se l'impresa è stata approvata o se il processo di data mining debba essere "pubblicizzato" come una tecnologia chiave per il gruppo di business.
- Se necessario, preparare presentazioni informative sul data mining nell'organizzazione.

Attività 3: descrivere la soluzione corrente

- Descrivere eventuali soluzioni già in uso per affrontare il problema di business.
- Descrivere i vantaggi e gli svantaggi della soluzione corrente. Indicare anche il livello di gradimento della soluzione in questione all'interno dell'azienda.

Definizione degli obiettivi di business

Questa è la fase in cui si definisce l'ambito di intervento specifico. In seguito alla ricerca condotta e alle riunioni cui si è preso parte, è necessario delineare un obiettivo primario concreto su cui convengano gli sponsor del progetto e le altre unità di business interessate dai risultati del data mining. Questo obiettivo si trasformerà, infine, da un'idea nebulosa come quella di "ridurre il tasso di abbandono dei clienti" in obiettivi di data mining specifici che guideranno le analisi successive.

Elenco delle attività

Prendere nota dei seguenti punti per poi incorporarli nel piano di progetto. È importante che gli obiettivi siano sempre realistici.

- Descrivere il problema che si desidera risolvere mediante il data mining.
- Indicare tutte le richieste di business nel modo più preciso possibile.
- Individuare eventuali altre esigenze di business, quali, ad esempio, la necessità di non perdere i clienti esistenti mentre si incrementano le opportunità di vendita incrociata.
- Indicare i vantaggi previsti in termini di business, ad esempio, la riduzione del 10% del tasso di abbandono dei clienti importanti.

Criteri per la valutazione dell'esito del business

Una volta definito l'obiettivo si deve però anche essere in grado di riconoscere quando lo si è raggiunto. Prima di procedere oltre, è importante definire quali siano i criteri di business con cui viene valutato l'esito del progetto di data mining. Tali criteri rientrano in due categorie:

- **Obiettivi.** Si tratta di criteri estremamente semplici, come la maggiore precisione dei controlli o una palese riduzione del tasso di abbandono dei clienti.
- **Soggettivi.** I criteri soggettivi, come l'individuazione di gruppi di trattamenti efficaci, sono più difficili da fissare con precisione, ma è possibile convenire su chi sia la persona incaricata della decisione finale.

Elenco delle attività

- Documentare, con la maggiore precisione possibile, i criteri per cui un progetto viene considerato riuscito.
- Verificare che ogni obiettivo di business abbia un criterio correlato per la valutazione dell'esito del progetto.
- Allineare le posizioni di coloro che si occuperanno di valutare se l'operazione abbia avuto o meno esito positivo. Se possibile, prendere nota delle loro aspettative.

Valutazione della situazione

Dopo aver definito chiaramente l'obiettivo, è opportuno valutare la situazione di partenza. In questa fase occorre porsi le seguenti domande:

- Che tipo di dati è possibile analizzare?
- Si ha a disposizione il personale necessario per completare il progetto?
- Quali sono i fattori di rischio principali del progetto?
- È stato ideato un piano di emergenza per ogni rischio?

Esempio di vendita al dettaglio in linea: valutazione della situazione

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Questo è il primo tentativo di Web-mining del rivenditore in linea di dispositivi elettronici e la società ha quindi deciso di consultare un esperto di data mining per agevolare l'operazione. Uno dei primi compiti che il consulente si trova ad affrontare è quello di valutare le risorse a disposizione della società per il data mining.

Personale. La società è provvista di personale interno esperto di gestione dei file di registro del server e di database di prodotti e acquisti, ma con poche conoscenze di data warehousing e pulitura dei dati per l'analisi. Quindi, può essere opportuno consultare anche un esperto di database. Poiché la società auspica che i risultati dello studio diventino parte integrante di un processo continuo di Web-mining, la dirigenza deve valutare anche se le posizioni create durante il progetto corrente debbano essere o meno permanenti.

Dati. Poiché si tratta di una società consolidata, sono disponibili molti dati di acquisto e log Web da analizzare. In realtà, per questo studio iniziale, la società restringerà l'analisi ai clienti che "hanno effettuato la registrazione" sul sito. In caso di esito positivo, il programma verrà esteso.

Rischi. A parte i costi dei consulenti e il tempo speso dai dipendenti per lo studio, l'impresa non presenta grandi probabilità di rischio immediato. Tuttavia, il fattore tempo è sempre importante, quindi il progetto iniziale viene pianificato per un solo trimestre finanziario.

Inoltre, poiché la disponibilità finanziaria è ridotta, è assolutamente indispensabile che lo studio rientri nei costi preventivati. Se esiste la possibilità che anche una sola di queste due condizioni non sia rispettata, i responsabili di business hanno suggerito di ridurre l'ambito del progetto.

Inventario delle risorse

Redigere un inventario accurato delle risorse è un passaggio indispensabile. Con un esame attento delle risorse hardware, delle sorgenti dati e del personale disponibili è possibile risparmiare tempo ed evitare problemi.

Attività 1: verificare le risorse hardware

- Quali sono i componenti hardware necessari?

Attività 2: identificare le sorgenti dati e gli archivi di informazioni

- Quali sono le sorgenti dati disponibili per il data mining? Prendere nota dei tipi di dati e dei relativi formati.
- Come sono archiviati i dati? È possibile accedere in tempo reale a data warehouse o database operativi?
- Si intende acquistare dati esterni, come informazioni demografiche?
- Esistono problemi di protezione che impediscono l'accesso ai dati richiesti?

Attività 3: identificare le risorse di personale

- È possibile avvalersi di esperti di business e dei dati?
- Sono stati identificati amministratori di database e altro personale di supporto, di cui potrebbe risultare necessario l'intervento?

Dopo essersi posti queste domande, includere un elenco di contatti e risorse per il report della fase.

Requisiti, presupposti e vincoli

Una valutazione onesta degli ostacoli al progetto aumenta le probabilità di riuscita. Rendere le proprie preoccupazioni quanto più esplicite possibile consente di evitare problemi futuri.

Attività 1: individuare i requisiti

Il requisito fondamentale è l'obiettivo di business di cui si è discusso in precedenza, ma occorre considerare quanto segue:

- Esistono vincoli legali o concernenti la sicurezza rispetto ai dati o ai risultati del progetto?
- Sono tutti d'accordo sui requisiti di pianificazione del progetto?
- Esistono requisiti per la distribuzione dei risultati, ad esempio, la pubblicazione sul Web o la lettura di punteggi in un database?

Attività 2: chiarire i presupposti

- Esistono fattori economici che potrebbero incidere sul progetto, ad esempio, compensi per le attività di consulenza o prodotti concorrenziali?
- Esistono presupposti di qualità dei dati?
- In che modo lo sponsor del progetto o il team di gestione desidera che gli siano presentati i risultati? In altre parole, è necessario rendere comprensibile il modello stesso o basta presentarne i risultati?

Attività 3: verificare i vincoli

- Sono disponibili tutte le password necessarie per l'accesso ai dati?
- Sono stati verificati tutti i vincoli legali relativi all'utilizzo dei dati?
- Nel budget del progetto sono stati rispettati tutti i vincoli finanziari?

Rischi e imprevisti

È consigliabile valutare i possibili rischi del progetto. Tra i tipi di rischi sono inclusi i seguenti:

- Pianificazione. Cosa accade se il progetto richiede più tempo del previsto?
- Finanziari. Cosa accade se lo sponsor del progetto si trova a dover affrontare problemi di budget?
- Dati. Cosa accade se i dati sono di qualità scadente o non esaustivi?
- Risultati. Cosa accade se i risultati iniziali sono meno significativi del previsto?

Dopo aver considerato i diversi rischi, definire un piano di emergenza per evitare una conclusione disastrosa.

Elenco delle attività

- Documentare ogni possibile rischio.

- Documentare un piano di emergenza per ogni rischio.

Terminologia

Per garantire che i team di business e di data mining "parlino la stessa lingua", è opportuno valutare l'ipotesi di compilare un glossario di termini tecnici e gergali che necessitano di un chiarimento. Ad esempio, se il termine "tasso di abbandono" ha per l'azienda un significato particolare e unico, è consigliabile renderlo esplicito a vantaggio dell'intero team. Allo stesso modo, il team può trarre beneficio dal chiarimento dell'uso di un grafico dei profitti.

Elenco delle attività

- Stendere un elenco dei termini o delle espressioni gergali non chiari per i membri del team. Includere sia la terminologia di business e che quella di data mining.
- Valutare l'ipotesi di pubblicare l'elenco sull'Intranet o in un'altra documentazione di progetto.

Analisi dei costi/benefici

La domanda cui, in questa fase, si fornisce una risposta è: **Quale sarà l'utile netto?** Nella valutazione finale è importante includere il confronto tra i costi del progetto e i potenziali vantaggi offerti da un esito positivo dello stesso.

Elenco delle attività

Includere nell'analisi i costi previsti per le seguenti operazioni:

- Raccolta dei dati ed eventuale acquisizione di dati esterni
- Distribuzione dei risultati
- Costi operativi

Quindi, prendere in considerazione i vantaggi di quanto di seguito riportato:

- Raggiungimento dell'obiettivo primario
- Informazioni aggiuntive ottenute dall'esplorazione dati
- Possibili vantaggi derivanti da una migliore comprensione dei dati

Definizione degli obiettivi di data mining

Dopo aver chiarito l'obiettivo di business, è necessario convertirlo in una realtà di data mining. Ad esempio, se lo scopo di business è "ridurre il tasso di abbandono", l'obiettivo del data mining includerà:

- Identificazione dei clienti importanti in base ai recenti dati di acquisto
- Creazione di un modello che si avvalga dei dati disponibili sui clienti per prevedere la probabilità di abbandono di ognuno di essi
- Assegnazione a ogni cliente di un punteggio basato sulla propensione all'abbandono e sul valore del cliente

Questi obiettivi di data mining, se raggiunti, possono poi essere utilizzati dall'azienda per ridurre il tasso di abbandono tra i clienti più importanti.

Come risulta evidente, il business e la tecnologia devono operare in sinergia per garantire un processo di data mining efficace. Nelle sezioni seguenti verranno forniti suggerimenti specifici per la definizione degli obiettivi di data mining.

Obiettivi del data mining

Quando si collabora con analisti di dati e di business per la definizione di una soluzione tecnica al problema di business, è importante avere sempre come obiettivo la concretezza.

Elenco delle attività

- Descrivere il **tipo** di problema di data mining, quale, ad esempio, il raggruppamento, la previsione o la classificazione.
- Documentare gli obiettivi tecnici utilizzando specifiche unità di tempo, ad esempio previsioni con una validità di tre mesi.
- Se possibile, fornire i numeri reali dei risultati desiderati, ad esempio la produzione di punteggi del tasso di abbandono per l'80% dei clienti esistenti.

Esempio di vendita al dettaglio in linea: obiettivi di data mining

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Con l'aiuto del consulente di data mining, il rivenditore in linea è stato in grado di tradurre gli obiettivi di business della società in termini di data mining. Gli obiettivi che lo studio iniziale deve raggiungere per il trimestre in corso sono i seguenti:

- Utilizzare le informazioni disponibili sui precedenti acquisti per generare un modello che colleghi gli articoli "correlati". Quando gli utenti osservano la descrizione di un articolo, fornire collegamenti ad altri articoli del gruppo correlato (**analisi market basket**).
- Utilizzare i log Web per determinare cosa cercano i diversi clienti e riprogettare quindi il sito per evidenziare tali articoli. Ogni "tipo" di cliente, con le caratteristiche peculiari, visualizzerà una pagina principale del sito diversa (**creazione di profili**).
- Utilizzare i log Web per tentare di prevedere quale sarà la mossa successiva di un utente sapendo da quale pagina proviene e quali pagine ha visitato (**analisi delle sequenze**).

Criteri per la valutazione dell'esito positivo del data mining

L'esito del processo deve essere definito anche in termini tecnici, per consentire una corretta gestione delle operazioni di data mining. Utilizzare gli obiettivi di data mining definiti in precedenza per formulare benchmark di riuscita del progetto. IBM SPSS Modeler fornisce strumenti, quali i nodi Valutazione e Analisi, che consentono di analizzare la precisione e la validità dei risultati.

Elenco delle attività

- Descrivere i metodi per la valutazione del modello, ad esempio, precisione, performance e così via.
- Definire i benchmark per la valutazione dell'esito del processo. Fornire valori specifici.
- Definire i sistemi di valutazione soggettivi nel modo più accurato possibile e stabilire chi deciderà dell'esito del processo.
- Decidere se la riuscita distribuzione del modello rientra tra i criteri di valutazione dell'esito del data mining. Iniziare a pianificare la distribuzione.

Produzione di un piano di progetto

A questo punto, si è pronti a elaborare un piano per il progetto di data mining. Le domande che ci si è posti fino ad ora e gli obiettivi di business e di data mining formulati costituiranno le basi di questa guida orientativa.

Stesura del piano di progetto

Il piano di progetto è il documento principale per tutte le operazioni di data mining. Se il documento viene redatto correttamente, è in grado di informare tutte le persone coinvolte nel progetto circa gli obiettivi, le risorse, i rischi e la pianificazione delle diverse fasi del data mining. Può essere opportuno pubblicare il piano, insieme alla documentazione raccolta in questa fase, nell'Intranet aziendale.

Elenco delle attività

Quando si redige il piano, verificare che si sia data risposta alle domande seguenti:

- Si è discusso delle attività del progetto e del piano proposto con tutte le persone coinvolte?

- Sono incluse previsioni temporali per tutte le fasi o le attività?
- Sono state incluse le attività e le risorse necessarie per la distribuzione dei risultati o della soluzione di business?
- Nel piano sono stati evidenziati i punti decisionali e le richieste di verifica?
- Sono state contrassegnate le fasi in cui si verificano, in genere, più iterazioni, come nella modellazione?

Esempio di piano di progetto

La tabella sottostante illustra il piano generale dello studio.

Tabella 1. Esempio di piano generale del progetto

Fase	Ora	Risorse	Rischi
Business Understanding	1 settimana	Tutti gli analisti	Cambiamento economico
Data Understanding	3 settimane	Tutti gli analisti	Problemi concernenti dati e tecnologie
Data Preparation	5 settimane	Consulente di data mining, analista di database (intervento parziale)	Problemi concernenti dati e tecnologie
Modellazione	2 settimane	Consulente di data mining, analista di database (intervento parziale)	Problemi concernenti le tecnologie, impossibilità di trovare un modello adeguato
Valutazione	1 settimana	Tutti gli analisti	Cambiamento economico, impossibilità di implementare i risultati
Distribuzione	1 settimana	Consulente di data mining, analista di database (intervento parziale)	Cambiamento economico, impossibilità di implementare i risultati

Valutazione di strumenti e tecniche

Dal momento che si è già scelto di servirsi di IBM SPSS Modeler per garantire la riuscita del data mining, è possibile utilizzare questo passaggio per ricercare le tecniche di data mining più appropriate alle diverse esigenze di business. IBM SPSS Modeler offre una vasta gamma di strumenti per ogni fase del processo. Per decidere quando utilizzare le diverse tecniche, consultare la sezione relativa alla modellazione della Guida in linea.

Considerazioni utili

Prima di passare ad analizzare i dati e a utilizzare IBM SPSS Modeler, accertarsi di aver fornito una risposta alle domande seguenti.

Da un punto di vista di business:

- Cosa si aspetta di ottenere l'azienda da questo progetto?
- Come si deciderà se il processo portato a termine ha avuto esito positivo?
- Sono disponibili il budget e le risorse necessari per raggiungere gli obiettivi prefissati?
- Si ha accesso a tutti i dati necessari per il progetto?

- Sono stati esaminati i rischi e gli imprevisti associati al progetto?
- I risultati dell'analisi costi/benefici rivelano la convenienza del progetto?

Dopo aver risposto alle domande precedenti, chiedersi se le risposte fornite sono state tradotte in un obiettivo di data mining.

Dal punto di vista del data mining:

- Qual è l'apporto specifico del data mining per il raggiungimento degli obiettivi di business?
- È già chiaro quali siano le tecniche di data mining più adeguate?
- Come è possibile sapere quando i risultati sono sufficientemente precisi ed efficaci? (*È stato impostato un sistema di valutazione dell'esito del data mining?*)
- Come saranno implementati i risultati della modellazione? Nel piano di progetto è stato incluso il processo di distribuzione?
- Il piano di progetto include tutte le fasi di CRISP-DM?
- Sono stati evidenziati rischi e fattori correlati?

Se la risposta alle domande riportate è affermativa, si è pronti a passare a un esame più attento dei dati.

Capitolo 3. Data Understanding

Panoramica sulla fase Data Understanding

La fase Data Understanding di CRISP-DM implica un esame più attento dei dati disponibili per il data mining. Questo passaggio è di fondamentale importanza per evitare problemi imprevisti durante la fase successiva di Data Preparation, che è in genere la parte più lunga di un progetto.

La comprensione dei dati implica l'accesso a essi e la successiva analisi mediante tabelle e grafici che possono essere organizzati in IBM SPSS Modeler utilizzando lo strumento per i progetti CRISP-DM. In questo modo è possibile determinare la qualità dei dati e descrivere i risultati di tali passaggi nella documentazione del progetto.

Raccolta di dati iniziali

A questo punto di CRISP-DM si è pronti ad accedere ai dati e a trasportarli in IBM SPSS Modeler. I dati provengono da diverse sorgenti, quali:

- **Dati esistenti.** Questa categoria include un'ampia gamma di dati: dati transazionali, dati di indagine, log Web e così via. Stabilire se i dati esistenti sono sufficienti a soddisfare le proprie esigenze.
- **Dati acquisiti.** Stabilire se l'azienda utilizza dati supplementari, come quelli demografici. In caso contrario, valutarne la necessità.
- **Dati aggiuntivi.** Se le sorgenti menzionate non soddisfano le proprie esigenze, potrebbe essere necessario condurre delle indagini o procedere a ulteriori registrazioni per integrare gli archivi di dati esistenti.

Elenco delle attività

Esaminare i dati in IBM SPSS Modeler e rispondere alle domande seguenti. Prendere nota delle proprie conclusioni. Per ulteriori informazioni, consultare la sezione “Creazione di un report di raccolta dei dati” a pagina 14.

- Quali attributi (colonne) del database sembrano più promettenti?
- Quali attributi sembrano irrilevanti e possono essere esclusi?
- I dati a propria disposizione sono sufficienti a tracciare conclusioni generalizzabili o a effettuare previsioni accurate?
- Sono presenti troppi attributi per il metodo di modellazione scelto?
- Si intende unire diverse sorgenti dati? In questo caso, sono presenti aree che potrebbero causare problemi durante l'unione?
- Sì è stabilito come gestire i valori mancanti in ognuna delle sorgenti dati?

Esempio di vendita al dettaglio in linea: raccolta di dati iniziali

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Il rivenditore in linea di questo esempio utilizza diverse sorgenti dati importanti, che includono le seguenti:

Log Web. I log degli accessi non elaborati contengono tutte le informazioni sul modo in cui i clienti esplorano il sito Web. Durante il processo di preparazione dei dati è necessario rimuovere dai log Web tutti i riferimenti ai file di immagine e altre voci non informative.

Dati sugli acquisti. Quando un cliente invia un ordine, tutte le informazioni pertinenti all'ordine vengono salvate. Gli ordini del database degli acquisti devono essere mappati alle sessioni corrispondenti dei log Web.

Database dei prodotti. Gli attributi dei prodotti possono essere utili per identificare prodotti "correlati". Le informazioni sui prodotti devono essere mappate agli ordini corrispondenti.

Database dei clienti. Questo database contiene informazioni aggiuntive, raccolte dai clienti registrati. I record non sono affatto completi, poiché molti clienti non compilano i questionari. Le informazioni sui clienti devono essere mappate agli acquisti e alle sessioni corrispondenti dei log Web.

Al momento, l'azienda non intende acquistare database esterni né spendere denaro per condurre indagini, poiché i suoi analisti sono occupati a gestire i dati attualmente a disposizione. A un certo punto, però, potrebbe essere opportuno implementare estesamente i risultati del data mining e in tal caso sarebbe utile acquisire dati demografici aggiuntivi per i clienti non registrati. Le informazioni demografiche possono, inoltre, essere utilizzate per individuare le differenze tra la base di clienti del rivenditore e la media degli acquirenti di prodotti sul Web.

Creazione di un report di raccolta dei dati

Con il materiale acquisito nel passaggio precedente è possibile creare un report di raccolta dei dati. Una volta completato, il report può essere aggiunto al sito Web del progetto o distribuito al team. Può anche essere combinato con i report che saranno preparati nei passaggi successivi, di descrizione dei dati, esplorazione e verifica della qualità. Questi report costituiranno una guida per la fase Data Preparation.

Descrizione dei dati

Esistono diversi modi per descrivere i dati, ma la maggior parte delle descrizioni si focalizza sulla quantità e la qualità: quanti dati sono disponibili e in che condizione sono. Di seguito sono elencate alcune caratteristiche chiave da tenere presenti quando si descrivono i dati.

- **Quantità di dati.** Per la maggior parte delle tecniche di modellazione esistono vantaggi e svantaggi associati alla dimensione dei dati. Insiemi di dati di grandi dimensioni possono produrre modelli più accurati, ma possono anche allungare i tempi di elaborazione. Valutare se è possibile utilizzare un sottoinsieme di dati. Quando si prendono appunti per il report finale, assicurarsi di includere statistiche sulle dimensioni per tutti gli insiemi di dati e, in fase di descrizione dei dati, ricordarsi di prendere in considerazione il numero dei record e i campi (attributi).
- **Tipi di valori.** I dati possono presentarsi in diversi formati: **numerico**, **categoriale** (stringa), o **booleano** (vero/falso). Se si presta attenzione al tipo di valore, è possibile evitare problemi durante la successiva modellazione.
- **Schemi di codifica.** Spesso i valori del database sono rappresentazioni di caratteristiche quali il sesso o il tipo di prodotto. Ad esempio, in un insieme di dati potrebbero essere utilizzate le lettere *M* e *F* per rappresentare il sesso *maschile* e *femminile*, mentre in un altro insieme potrebbero essere utilizzati i valori numerici *1* and *2*. Si annoti ogni schema contraddittorio nel report dei dati.

Con le conoscenze acquisite, è ora possibile creare il report di descrizione dei dati e condividere le conclusioni cui si è giunti con una più ampia platea di interlocutori.

Esempio di vendita al dettaglio in linea: descrizione dei dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

In un'applicazione di Web-mining devono essere elaborati molti record e attributi. Anche se il rivenditore in linea che conduce il progetto di data mining ha limitato lo studio iniziale a circa 30.000 clienti registrati sul sito, sono comunque presenti milioni di record nei log Web.

Quasi tutti i tipi di valori presenti in queste sorgenti dati sono simbolici, a prescindere dal fatto che rappresentino date e ore, pagine Web visitate o risposte a domande a scelta multipla del questionario di registrazione. Alcune delle variabili verranno utilizzate per creare nuove variabili di tipo numerico, ad esempio, il numero di pagine Web visitate e il tempo trascorso sul sito Web. Le poche variabili numeriche

esistenti nelle sorgenti dati includono il numero di ogni prodotto ordinato, l'importo dell'acquisto e le specifiche di peso e dimensione dei prodotti ricavate dal database dei prodotti.

La sovrapposizione degli schemi di codifica delle diverse sorgenti dati è estremamente ridotta, poiché tali sorgenti contengono attributi molti diversi. Le sole variabili che si sovrappongono sono quelle "chiave", quali gli ID dei clienti e i codici dei prodotti. Queste variabili devono avere schemi di codifica identici da una sorgente dati all'altra, altrimenti, non sarà possibile unire le diverse sorgenti. Per ricodificare questi campi chiave per l'unione, sono necessarie alcune ulteriori operazioni di preparazione dei dati.

Creazione di un report di descrizione dei dati

Per procedere in maniera efficace allo svolgimento del progetto di data mining, considerare l'importanza di creare un accurato report di descrizione dei dati utilizzando i seguenti criteri di valutazione:

Quantità dei dati

- Formato dei dati
- Metodo utilizzato per acquisire i dati, ad esempio, ODBC.
- Dimensioni del database, in termini di righe e colonne.

Qualità dei dati

- Caratteristiche dei dati rilevanti per la domanda di business.
- Tipi di dati disponibili (simbolici, numerici e così via).
- Statistiche di base calcolate per gli attributi chiave. Apporto di tali statistiche alla comprensione della domanda di business.
- Possibilità di definire priorità tra gli attributi pertinenti. Disponibilità degli analisti di business a fornire ulteriori informazioni esplicative, nel caso non sia possibile definire priorità.

Analisi dei dati

Utilizzare questa fase di CRISP-DM per analizzare i dati avvalendosi di tabelle, grafici e altri strumenti di visualizzazione disponibili in IBM SPSS Modeler. Tali analisi possono agevolare il raggiungimento dell'obiettivo di data mining identificato durante la fase Business Understanding. Consentono inoltre di formulare ipotesi e approntare attività di trasformazione dati che verranno eseguite durante la fase Data Preparation.

Esempio di vendita al dettaglio in linea: analisi dei dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Sebbene si consigli di eseguire un'analisi iniziale in questa fase del processo, l'esplorazione dati è un'operazione complessa, se non impossibile, su log Web non elaborati, come ha potuto rilevare il rivenditore in linea. In genere, perché siano significativamente analizzabili, i dati dei log Web devono prima essere elaborati nella fase Data Preparation. Questa presa di distanza da CRISP-DM sottolinea il fatto che il processo può e deve essere adattato alle diverse esigenze di data mining. CRISP-DM è uno strumento ciclico e i data miner si muovono in genere avanti e indietro tra le fasi.

Mentre i log Web devono essere elaborati prima dell'esplorazione, le altre sorgenti dati disponibili risultano più facilmente esplorabili. L'uso del database degli acquisti per l'esplorazione rivela interessanti informazioni sui clienti, ad esempio quanto spendono, quanti articoli comprano per ogni acquisto e da dove provengono. I riepiloghi del database dei clienti mostreranno, invece, la distribuzione delle risposte alle domande del questionario di registrazione.

L'esplorazione è utile anche per individuare errori nei dati. Mentre la maggior parte delle sorgenti dati viene generata automaticamente, le informazioni del database dei prodotti vengono immesse manualmente. Alcuni rapidi riepiloghi delle dimensioni dei prodotti in elenco aiuteranno a individuare errori di battitura, ad esempio, monitor da 119 pollici, invece che da 19 pollici.

Creazione di un report di esplorazione dati

Quando si creano grafici e si eseguono statistiche sui dati disponibili, iniziare a formulare ipotesi sul modo in cui tali dati possano fornire una risposta agli obiettivi tecnici e di business.

Elenco delle attività

Prendere nota delle proprie conclusioni per includerle nel report di esplorazione dati. Rispondere alle domande seguenti:

- Che tipo di ipotesi sono state formulate sui dati?
- Quali attributi sembrano promettenti per un'ulteriore analisi?
- Le esplorazioni hanno evidenziato nuove caratteristiche dei dati?
- In che modo tali esplorazioni hanno modificato l'ipotesi iniziale?
- Si è in grado di identificare particolari sottoinsiemi di dati da utilizzare in seguito?
- Esaminare nuovamente gli obiettivi di data mining. L'esplorazione li ha in qualche modo modificati?

Verifica della qualità dei dati

Accade raramente che i dati siano perfetti. La maggior parte contiene errori di codifica, valori mancanti e altri tipi di incoerenze che talvolta rendono l'analisi piuttosto complessa. Uno dei modi per evitare potenziali insidie è quello di condurre un'approfondita analisi della qualità dei dati disponibili prima della modellazione.

Gli strumenti di reporting disponibili in IBM SPSS Modeler (ad esempio i nodi Esplora, Tabella e altri nodi di output) possono semplificare la ricerca delle seguenti tipologie di problemi:

- **Dati mancanti**: includono valori vuoti o codificati come una non risposta (ad esempio, *$null$*, *?* o *999*).
- **Errori nei dati**: si tratta in genere di errori tipografici introdotti durante l'immissione dei dati.
- **Errori di misurazione**: includono dati immessi correttamente ma basati su una schema di misurazione non corretto.
- **Incoerenze di codifica**: includono di norma unità di misura non standard o incoerenze nei valori, quali l'uso di entrambi gli elementi *M* e *maschile* per indicare il sesso.
- **Metadati non validi**: includono discordanze tra il significato apparente di un campo e il significato implicito nel nome o nella definizione del campo.

Prendere nota dei problemi concernenti la qualità. Per ulteriori informazioni, consultare la sezione “Creazione di un report sulla qualità dei dati” a pagina 17.

Esempio di vendita al dettaglio in linea: verifica della qualità dei dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

La verifica della qualità dei dati viene spesso eseguita nel corso dei processi di descrizione ed esplorazione. Tra i problemi riscontrati dal rivenditore in linea sono inclusi i seguenti:

Dati mancanti. I dati mancanti noti includono questionari non completati da utenti registrati. Senza le informazioni aggiuntive fornite dal questionario, questi clienti potrebbero dover essere esclusi da alcuni dei modelli successivi.

Dati errati. La maggior parte delle sorgenti dati viene generata automaticamente, quindi non comporta grossi problemi. Durante il processo di esplorazione possono essere individuati errori tipografici nel database dei prodotti.

Errori di misurazione. La principale fonte potenziale di errori di misurazione è il questionario. Se include elementi mal pubblicizzati o non correttamente espressi, il questionario potrebbe non fornire le risposte che il rivenditore spera di ottenere. Inoltre, durante il processo di esplorazione, è importante prestare particolare attenzione agli elementi con un'inusuale distribuzione di risposte.

Creazione di un report sulla qualità dei dati

Basandosi sui risultati dei processi di esplorazione e di verifica della qualità dei dati è possibile, a questo punto, passare a preparare un report che farà da guida per la successiva fase di CRISP-DM. Per ulteriori informazioni, consultare la sezione “Verifica della qualità dei dati” a pagina 16.

Elenco delle attività

Come si è detto in precedenza, esistono diversi tipi di problemi concernenti la qualità dei dati. Prima di proseguire oltre, valutare i seguenti problemi di qualità e ipotizzare una soluzione. Documentare tutte le risposte nel report sulla qualità dei dati.

- Sono stati individuati attributi mancanti e campi vuoti? In questo caso, esiste un significato dietro tali valori mancanti?
- Sono presenti incoerenze ortografiche che possono causare problemi nelle successive unioni o trasformazioni?
- Sono state esaminate le deviazioni per stabilire se si trattava di "rumore" o di fenomeni degni di un'ulteriore analisi?
- È stato eseguito un controllo di plausibilità dei valori? Prendere nota dei conflitti apparenti, ad esempio, adolescenti con livelli di reddito elevati.
- È stata valutata la possibilità di escludere dati che non hanno alcun impatto sulle ipotesi formulate?
- I dati sono archiviati in file flat? In questo caso, i delimitatori tra i file sono coerenti? Ogni record contiene lo stesso numero di campi?

Considerazioni utili

Prima di preparare i dati per la modellazione in IBM SPSS Modeler, valutare i punti seguenti:

Qual è il livello di comprensione dei dati?

- Tutte le sorgenti dati sono state chiaramente identificate e consultate? Si è a conoscenza di problemi o restrizioni?
- Sono stati identificati attributi chiave dai dati disponibili?
- Questi attributi agevolano in qualche modo la formulazione di ipotesi?
- È stata valutata la dimensione di tutte le sorgenti dati?
- È possibile utilizzare un sottoinsieme di dati nei casi appropriati?
- Sono state calcolate statistiche di base per ciascun attributo di interesse? Sono emerse informazioni significative?
- Sono stati utilizzati grafici esplicativi per migliorare la comprensione degli attributi chiave? Le ulteriori informazioni, acquisite in questo modo, hanno costretto a riformulare le ipotesi?
- Quali sono i problemi di qualità dei dati per il progetto? È stato ideato un piano per risolvere tali problemi?
- I passaggi di preparazione dei dati sono chiari? Ad esempio, si è compreso quali sorgenti dati unire e quali attributi filtrare o selezionare?

A questo punto, avendo completa comprensione degli obiettivi di business e dei dati a disposizione, è giunto il momento di utilizzare IBM SPSS Modeler per preparare i dati per la modellazione.

Capitolo 4. Data Preparation

Panoramica sulla fase Data Preparation

La preparazione dei dati è uno degli aspetti più importanti del data mining e richiede spesso tempi piuttosto lunghi. È stato calcolato, infatti, che tale operazione richiede in genere dal 50 al 70% del tempo e dell'energia spesi in un progetto. Dedicare l'adeguata energia alle fasi preliminari di Business Understanding e Data Understanding consente di ridurre l'overhead, ma per la preparazione e l'assemblaggio dei dati saranno comunque necessari tempo e fatica.

A seconda degli obiettivi dell'azienda, la preparazione dei dati implica le seguenti attività:

- Unione di insiemi e/o record di dati
- Selezione di un sottoinsieme di dati campione
- Aggregazione di record
- Derivazione di nuovi attributi
- Ordinamento dei dati per la modellazione
- Rimozione o sostituzione di valori vuoti o mancanti
- Suddivisione in insiemi di dati di addestramento e test

Selezione di dati

Basandosi sulla raccolta di dati iniziali, condotta nella precendente fase di CRISP-DM, è possibile iniziare a selezionare i dati pertinenti agli obiettivi di data mining. In genere, esistono due modi di selezione dei dati:

- La **selezione di elementi (righe)** implica la necessità di prendere decisioni su quali account, prodotti o clienti includere.
- La **selezione di attributi o caratteristiche (colonne)** implica, invece, la necessità di prendere decisioni relative all'uso di caratteristiche quali l'importo delle transazioni o il reddito familiare.

Esempio di vendita al dettaglio in linea: selezione dei dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Molte delle decisioni del rivenditore in linea circa i dati da selezionare sono state prese nelle prime fasi del processo di data mining.

Selezione di elementi. Lo studio iniziale sarà limitato a circa 30.000 clienti registrati sul sito. Sarà quindi necessario impostare dei filtri per escludere acquisti e log Web di clienti non registrati. Dovranno poi essere definiti altri filtri per rimuovere le chiamate a file di immagine e altre voci non informative dei log Web.

Selezione di attributi. Il database degli acquisti conterrà informazioni riservate sui clienti del rivenditore in linea, quindi è importante filtrare attributi quali il nome del cliente, l'indirizzo, il numero di telefono e i codici di carta di credito.

Inclusione o esclusione dei dati

Quando si decide quali sottoinsiemi di dati includere o escludere, indicare la logica che ha guidato la decisione.

Domande da tenere presenti

- Che importanza ha un determinato attributo nel raggiungimento degli obiettivi di data mining?
- La qualità di un particolare insieme di dati o di un determinato attributo preclude la validità dei risultati?
- È possibile recuperare tali dati?
- Esistono vincoli relativi all'uso di campi specifici quali *sesso* o *gruppo etnico*?

Le decisioni prese si discostano dalle ipotesi formulate nella fase Data Understanding? In questo caso, indicare il motivo di tali decisioni nel report del progetto.

Clean Data

La pulitura dei dati implica una verifica attenta dei problemi presenti nei dati che si è scelto di includere nell'analisi. Esistono diversi metodi per pulire i dati utilizzando i nodi Record e Operazioni sui campi di IBM SPSS Modeler.

Tabella 2. Pulitura dati

Problema dei dati	**Possibile soluzione**
Dati mancanti	Escludere righe o caratteristiche oppure riempire i vuoti con un valore stimato.
Dati errati	Utilizzare la logica per individuare manualmente gli errori e correggerli oppure escludere caratteristiche.
Incoerenze di codifica	Scegliere un singolo schema di codifica, quindi convertire e sostituire i valori.
Metadati mancanti o non validi	Esaminare manualmente i campi sospetti e individuarne il significato corretto.

Il report sulla qualità dei dati, preparato durante la fase Data Understanding, contiene dettagli sui tipi di problemi specifici dei dati scelti. È quindi possibile avvalersene come punto di partenza per la manipolazione dei dati in IBM SPSS Modeler.

Esempio di vendita al dettaglio in linea: pulitura dei dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Il rivenditore in linea utilizza il processo di pulitura per affrontare i problemi rilevati nel report sulla qualità dei dati.

Dati mancanti. I clienti che non hanno completato il questionario in linea potrebbero, in seguito, dover essere esclusi da alcuni dei modelli. È possibile chiedere nuovamente a tali clienti di compilare il questionario, ma questo implicherebbe tempo e denaro, risorse che il rivenditore in linea non può permettersi di sprecare. Ciò che, però, il rivenditore può fare è modellare le differenze di acquisto tra i clienti che hanno risposto e quelli che non hanno risposto alle domande del questionario. Se questi due insiemi di clienti hanno abitudini di acquisto simili, i questionari mancanti perdono rilevanza.

Dati errati. In questa fase è possibile correggere gli errori riscontrati durante il processo di esplorazione. Nella maggior parte dei casi, sul sito Web l'immissione di dati corretti viene forzata prima che un cliente invii una pagina al database back-end.

Errori di misurazione. Se nel questionario sono presenti elementi non correttamente espressi, la qualità dei dati può risultare significativamente compromessa. Si tratta di un problema di difficile soluzione, esattamente come quello dei questionari mancanti, perché potrebbe non esserci tempo né denaro per raccogliere risposte a una nuova domanda sostitutiva. Per gli elementi problematici, la soluzione migliore è ritornare al processo di selezione e filtrare tali elementi per escluderli dalle ulteriori analisi.

Creazione di un report di pulitura dei dati

La creazione di un report relativo alle operazioni di pulitura dei dati è un passaggio fondamentale per tenere traccia delle alterazioni ai dati. I futuri progetti di data mining trarranno vantaggio dalla disponibilità immediata di informazioni dettagliate sulle operazioni eseguite.

Elenco delle attività

Quando si crea un report di questo tipo è buona norma tenere presenti le seguenti domande:

- Quale tipo di rumore è stato generato nei dati?
- Quali sono i metodi utilizzabili per eliminare il rumore? Quali tecniche hanno avuto esito positivo?
- Sono presenti casi o attributi che potrebbero non essere recuperati? Prendere nota dei dati esclusi a causa del rumore.

Creazione di nuovi dati

Accade spesso di dover creare nuovi dati. Ad esempio, può essere utile creare una nuova colonna che contrassegni l'acquisto di una garanzia estesa per ogni transazione. Questo nuovo campo, *purchased_warranty*, può essere generato senza alcuna difficoltà utilizzando un nodo Crea flag in IBM SPSS Modeler.

Esistono due metodi per creare nuovi dati:

- Derivazione di attributi (colonne o caratteristiche)
- Generazione di record (righe)

IBM SPSS Modeler prevede diversi metodi per creare dati utilizzando i nodi Record e Operazioni di campo.

Esempio di vendita al dettaglio in linea: creazione di dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Con l'elaborazione dei log Web è possibile creare molti nuovi attributi. Per gli eventi registrati nei file di registro, il rivenditore desidera creare timestamp, identificare visitatori e sessioni e annotare la pagina cui si è avuto accesso e il tipo di attività che l'evento rappresenta. Alcune di queste variabili verranno utilizzate per creare ulteriori attributi, come il tempo trascorso tra gli eventi in una sessione.

Gli attributi aggiuntivi possono essere creati dall'unione o dalla riorganizzazione di altri dati. Ad esempio, se i log Web strutturati con un evento per riga vengono elaborati in modo che ogni riga corrisponda a una sessione, verranno creati nuovi attributi che indicano il numero totale di operazioni, il tempo complessivo trascorso sul sito e il numero totale di acquisti effettuati durante la sessione. Se i log Web vengono uniti al database dei clienti, in modo che ogni riga corrisponda a un cliente, verranno creati nuovi attributi che indicano il numero di sessioni, il numero totale di operazioni, il tempo complessivo trascorso sul sito e il numero totale di acquisti effettuati da ogni cliente.

Dopo la creazione dei dati, il rivenditore in linea esegue un processo di esplorazione per assicurarsi che l'operazione di creazione sia stata eseguita correttamente.

Derivazione di attributi

In IBM SPSS Modeler è possibile utilizzare i nodi Operazioni di campo seguenti per derivare nuovi attributi:

- Il **nodo Ricava** consente di creare nuovi campi derivati da campi esistenti.
- Il **nodo Crea flag** consente di creare un campo flag.

Elenco delle attività

- Quando si derivano attributi tenere presenti i requisiti dei dati per la modellazione. Se, ad esempio, l'algoritmo di modellazione richiede un particolare tipo di dati, come quelli numerici, eseguire le trasformazioni necessarie.

- Valutare se i dati debbano essere normalizzati prima della modellazione.
- Stabilire se gli attributi mancanti possano essere creati mediante aggregazione, ripartizione proporzionale o induzione.
- In base alle proprie conoscenze, stabilire se esistono elementi importanti, come il periodo di tempo trascorso sul sito Web, che possano essere derivati dai campi esistenti.

Integrazione dei dati

Capita spesso di disporre di più sorgenti dati per lo stesso insieme di domande di business. Ad esempio, se si ha accesso ai dati relativi ai prestiti ipotecari e a dati demografici acquisiti dall'esterno per lo stesso gruppo di clienti e se tali insiemi di dati contengono lo stesso identificatore univoco, quale, ad esempio, il codice fiscale, è possibile unirli in IBM SPSS Modeler mediante questo campo chiave.

Esistono due metodi di base di integrazione dei dati:

- L'**unione** implica la fusione di due insiemi di dati con record simili ma attributi diversi. I dati verranno uniti utilizzando lo stesso identificatore chiave per ogni record, ad esempio l'ID del cliente. I dati risultanti prevederanno un maggior numero di colonne o caratteristiche.
- L'**accodamento** implica, invece, l'integrazione di due insiemi di dati con attributi simili ma record diversi. I dati verranno integrati in base ai campi simili, ad esempio quelli relativi al nome del prodotto o alla lunghezza del contratto.

Esempio di vendita al dettaglio in linea: integrazione dei dati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Avendo a disposizione più sorgenti dati, esistono molti modi per integrare i dati:

- **Aggiunta di attributi di clienti e prodotti ai dati sugli eventi.** Per modellare gli eventi dei log Web utilizzando attributi di altri database, è necessario identificare correttamente qualsiasi ID cliente, numero di prodotto e numero di ordine di acquisto associato a ogni evento e unire gli attributi corrispondenti ai log Web elaborati. Si noti che il file unito replica le informazioni sui clienti e i prodotti ogni volta che un cliente o un prodotto viene associato a un evento.
- **Aggiunta delle informazioni di acquisto e dei log Web ai dati sui clienti.** Per modellare il valore di un cliente, è necessario selezionare le informazioni sugli acquisti e le sessioni dai database appropriati, raccoglierle e unirle al database dei clienti. Questo processo implica la creazione di nuovi attributi, come descritto nella procedura di creazione dei dati.

Dopo l'integrazione dei database, il rivenditore in linea esegue un processo di esplorazione per assicurarsi che l'operazione di unione dei dati sia stata effettuata correttamente.

Attività di integrazione

L'integrazione dei dati può diventare piuttosto complessa se non si è dedicato il tempo adeguato alla comprensione degli stessi. Riflettere sugli elementi e gli attributi più importanti per gli obiettivi di data mining e procedere quindi a integrare i dati.

Elenco delle attività

- Utilizzando i nodi Unione e Accodamento di IBM SPSS Modeler, integrare gli insiemi di dati considerati utili per la modellazione.
- Valutare l'opportunità di salvare l'output prodotto prima di procedere alla modellazione.
- Dopo l'unione, i dati possono essere semplificati **aggregando** i valori. Nell'aggregazione i nuovi valori vengono calcolati riepilogando le informazioni di più record e/o tabelle.
- Potrebbe anche essere necessario generare nuovi record, come, ad esempio, la detrazione media di diversi anni di dichiarazioni dei redditi combinate.

Formattazione dei dati

Come passaggio finale prima della creazione del modello, è utile verificare se determinate tecniche richiedano un particolare formato o ordine dei dati. Ad esempio, non è raro che un algoritmo sequenziale richieda che i dati vengano preordinati prima dell'esecuzione del modello. Anche se il modello è in grado di eseguire l'ordinamento automaticamente, l'utilizzo del nodo Ordina prima della modellazione consente di ridurre i tempi di elaborazione.

Elenco delle attività

Quando si formattano i dati tenere presenti le domande seguenti:

- Quali modelli si intende utilizzare?
- Questi modelli richiedono un particolare formato o ordine dei dati?

Se è consigliabile eseguire modifiche, gli strumenti di elaborazione di IBM SPSS Modeler consentono di effettuare le necessarie manipolazioni dei dati.

Considerazioni per la modellazione

Prima di creare modelli in IBM SPSS Modeler, accertarsi di aver risposto alle domande seguenti.

- I dati sono tutti accessibili da IBM SPSS Modeler?
- In base all'esplorazione iniziale e alla comprensione dei dati, è stato possibile selezionare sottoinsiemi di dati pertinenti?
- Sono stati efficamente puliti i dati o rimossi gli elementi non recuperabili? Indicare qualsiasi decisione nel report finale.
- I diversi insiemi di dati sono stati integrati correttamente? Sono stati riscontrati problemi di unione che dovrebbero essere documentati?
- Sono stati esaminati i requisiti degli strumenti di modellazione che si intende utilizzare?
- Sono presenti problemi di formattazione risolvibili prima della modellazione? Nella risposta alla domanda includere le attività di formattazione necessarie e quelle che potrebbero ridurre i tempi di modellazione.

Se si è in grado di fornire una risposta alle domande riportate, si può passare al punto cruciale del data mining, la modellazione.

Capitolo 5. Modellazione

Panoramica sulla fase di modellazione

È questo il momento in cui tutti gli sforzi cominciano a dare risultati. I dati che si è passato tanto tempo a preparare vengono trasferiti negli strumenti di analisi di IBM SPSS Modeler e i risultati cominciano a fare luce sul problema di business posto durante la fase Business Understanding.

La modellazione viene in genere eseguita in più iterazioni. Di norma, i data miner eseguono diversi modelli utilizzando i parametri di default, per poi regolare i parametri o tornare alla fase di Data Preparation per le manipolazioni richieste dal modello scelto. È raro che la domanda di data mining di un'azienda riceva una risposta soddisfacente con un solo modello e una sola esecuzione. Questo è ciò che rende il data mining così interessante: esistono molti modi per esaminare un determinato problema e IBM SPSS Modeler offre un'ampia gamma di strumenti utili allo scopo.

Selezione delle tecniche di modellazione

Anche se è già in qualche modo chiaro quali siano i tipi di modellazione più appropriati alle esigenze dell'azienda, è giunto il momento di prendere delle decisioni definitive sulle tecniche da utilizzare. La decisione circa il modello più appropriato si basa in genere sulle seguenti considerazioni:

- **Tipi di dati disponibili per il data mining.** Ad esempio, i campi di interesse sono categoriali (simbolici)?
- **Obiettivi del data mining.** Si desidera semplicemente comprendere meglio gli archivi di dati transazionali e svelare interessanti schemi di acquisto oppure, è necessario produrre un punteggio che indichi, ad esempio, la propensione a non restituire un prestito?
- **Requisiti di modellazione specifici.** Il modello richiede dati di un particolare tipo o di dimensioni specifiche? Si ha bisogno di un modello con risultati facilmente presentabili?

Per ulteriori informazioni sui tipi di modelli disponibili in IBM SPSS Modeler e i relativi requisiti, fare riferimento alla documentazione o alla Guida in linea di IBM SPSS Modeler.

Esempio di vendita al dettaglio in linea: tecniche di modellazione

Le tecniche di modellazione impiegate dal rivenditore in linea sono state scelte in base agli obiettivi di data mining dell'azienda:

Miglioramento delle indicazioni. Questa operazione implica, nel suo aspetto più immediato, il raggruppamento degli ordini di acquisto, per stabilire quali prodotti vengono più frequentemente acquistati insieme. Per ampliare i risultati è possibile aggiungere i dati dei clienti e anche i record delle visite. Per questo tipo di modellazione sono adeguate le tecniche di cluster rete a due fasi o di Kohonen. In seguito, è possibile creare profili dei cluster utilizzando un insieme di regole C5.0, per determinare quali sono le indicazioni più appropriate nei diversi punti della visita di un cliente.

Miglioramento dell'esplorazione del sito. Per il momento, il rivenditore in linea si concentrerà sull'identificazione delle pagine utilizzate di frequente ma che richiedono diversi passaggi per essere trovate. Questa operazione implica l'applicazione di un algoritmo sequenziale ai log Web, al fine di generare i "percorsi univoci" dei clienti attraverso il sito Web e, quindi, di ricercare specificamente le sessioni che presentano un numero elevato di visite per pagina senza o prima dell'esecuzione di qualsiasi operazione. In un secondo momento, nel corso di un'analisi più dettagliata, le tecniche di raggruppamento possono essere utilizzate per identificare diversi "tipi" di visite e visitatori e il contenuto del sito può essere organizzato e presentato in base al tipo individuato.

Scelta delle tecniche di modellazione appropriate

IBM SPSS Modeler propone numerose tecniche di modellazione. I data miner utilizzano spesso più di una tecnica per affrontare il problema da diverse angolazioni.

Elenco delle attività

Quando si decide quali modelli utilizzare, valutare se i seguenti fattori incidono sulle scelte:

- Il modello richiede che i dati siano suddivisi in set di test e di addestramento?
- Sono disponibili dati sufficienti a produrre risultati affidabili per un determinato modello?
- Il modello richiede un certo livello di qualità dei dati? È possibile raggiungere tale livello con i dati correnti?
- Il tipo di dati è appropriato per uno specifico modello? Nel caso non lo sia, è possibile eseguire le conversioni necessarie utilizzando nodi di manipolazione dei dati?

Per ulteriori informazioni sui tipi di modelli disponibili in IBM SPSS Modeler e i relativi requisiti, fare riferimento alla documentazione o alla Guida in linea di IBM SPSS Modeler.

Presupposti della modellazione

Quando si inizia a restringere il campo degli strumenti di modellazione selezionabili, prendere appunti sul processo decisionale. Documentare qualsiasi ipotesi sui dati e qualsiasi manipolazione degli stessi eseguita al fine di soddisfare i requisiti del modello.

Ad esempio, i nodi Regressione logistica e Rete neurale richiedono entrambi che i tipi di dati siano completamente **istanziati** (tipi di dati noti) prima dell'esecuzione. Questo significa che sarà necessario aggiungere un nodo Tipo al flusso ed eseguirlo, per passare in rassegna i dati prima di creare ed eseguire un modello. Allo stesso modo, i modelli predittivi, quale C5.0, possono trarre vantaggio dal ribilanciamento dei dati quando vengono previste regole per gli eventi rari. Quando si esegue questo tipo di previsione, è possibile ottenere risultati migliori inserendo un nodo bilanciamento nel flusso e introducendo nel modello il sottoinsieme più bilanciato.

Non tralasciare di documentare questo tipo di decisioni.

Generazione di un progetto di test

Prima di creare il modello occorre valutare attentamente come se ne verificheranno i risultati. La generazione di un progetto di test esaustivo include due parti:

- Descrizione dei criteri di "bontà" di un modello
- Definizione dei dati sui cui questi criteri verranno verificati

La **bontà** di un modello può essere misurata in diversi modi. Per i modelli supervisionati, quali C5.0 e C&R Tree, viene presa in considerazione la frequenza di errore di un particolare modello. Per i modelli non supervisionati, come le reti cluster Kohonen, la misurazione può includere criteri quali la semplicità di interpretazione e di distribuzione o i tempi di elaborazione necessari.

È importante ricordare che la creazione del modello è un processo iterativo. Questo significa che, prima di decidere quali modelli utilizzare e implementare, verranno verificati i risultati di diversi modelli.

Creazione di un progetto di test

Il progetto di test è una descrizione dei passaggi da eseguire per verificare i modelli prodotti. Poiché la modellazione è un processo iterativo, è importante sapere quando smettere di regolare i parametri e tentare con un altro metodo o modello.

Elenco delle attività

Quando si crea un progetto di test, tenere presenti le domande seguenti:

- Quali dati verranno utilizzati per verificare i modelli? I dati sono stati suddivisi in insiemi di addestramento e di test (approccio comunemente utilizzato nella modellazione)?
- Come è possibile valutare la riuscita dei modelli supervisionati (come C5.0)?
- Come è possibile valutare la riuscita dei modelli non supervisionati (come le reti cluster Kohonen)?
- Quante volte si è disposti a rieseguire un modello con impostazioni rettificate prima di provarne un altro?

Esempio di vendita al dettaglio in linea: progetto di test

Scenario di Web-mining in cui viene utilizzato CRISP-DM

I criteri in base ai quali vengono valutati i modelli dipendono dai modelli in questione e dagli obiettivi di data mining:

Miglioramento delle indicazioni. Fino a quando le indicazioni migliorate non vengono presentate ai clienti reali, non esiste un modo oggettivo di valutarle. Tuttavia, il rivenditore in linea può richiedere che le regole che generano le indicazioni siano sufficientemente semplici da essere coerenti in una prospettiva di business. Allo stesso tempo, però, tali regole devono essere sufficientemente complesse da generare diverse indicazioni per diversi clienti e sessioni.

Miglioramento dell'esplorazione del sito. Una volta individuate le pagine del sito Web visitate dai clienti, il rivenditore in linea può valutare oggettivamente la struttura del sito aggiornato in termini di semplicità di accesso alle pagine importanti. Tuttavia, come per le indicazioni, è difficile valutare in anticipo quale sarà la reazione dei clienti rispetto alla riorganizzazione del sito. Se sono disponibili tempo e denaro, potrebbe essere utile eseguire dei test di utilizzabilità.

Creazione dei modelli

A questo punto, si dovrebbe essere pronti a creare i modelli su cui si è tanto riflettuto. Occorre tempo e spazio per sperimentare diversi modelli prima di giungere a conclusioni definitive. La maggior dei data miner crea diversi modelli e ne confronta i risultati prima di implementarli o integrarli.

Per tenere traccia di quanto appreso sui diversi modelli, prendere nota delle impostazioni e dei dati utilizzati per ciascun modello. Questo consentirà di descrivere i propri risultati ad altri e di ripercorrere i passaggi eseguiti, se necessario. Al termine del processo di creazione di modelli, saranno disponibili tre gruppi di informazioni da utilizzare per le decisioni di data mining:

- Le **impostazioni dei parametri**, che includono gli appunti presi sui parametri che producono i risultati migliori.
- I **modelli** realmente prodotti.
- Le **descrizioni dei risultati dei modelli**, inclusi i problemi concernenti performance e dati che si sono verificati durante l'esecuzione del modello e l'esplorazione dei risultati.

Esempio di vendita al dettaglio in linea: creazione del modello

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Miglioramento delle indicazioni. Vengono prodotti raggruppamenti per diversi livelli di integrazione dei dati, partendo dal solo database degli acquisti per poi includere le informazioni correlate su clienti e sessioni. Per ogni livello di integrazione, vengono prodotti raggruppamenti con diverse impostazioni di parametri per gli algoritmi di rete a due fasi e Kohonen. Per ciascun raggruppamento, vengono generati alcuni insiemi di regole C5.0 con diverse impostazioni di parametri.

Miglioramento dell'esplorazione del sito. Il nodo Modelli Sequenza viene utilizzato per generare percorsi dei clienti. L'algoritmo consente di specificare un criterio minimo di supporto, utile per concentrarsi sui percorsi dei clienti più comuni. Per i parametri vengono provate diverse impostazioni.

Impostazioni dei parametri

La maggior parte delle tecniche di modellazione prevedono diversi parametri o impostazioni regolabili per controllare il processo di modellazione. Ad esempio, le strutture ad albero delle decisioni possono essere controllate regolandone la profondità, le suddivisioni e diverse altre impostazioni. In genere, la maggior parte delle persone crea un modello utilizzando prima le opzioni di default e quindi ridefinendo i parametri durante le successive sessioni.

Una volta determinati i parametri che producono i risultati più accurati, salvare il flusso e i nodi del modello generato. Inoltre, prendere nota delle impostazioni ottimali può essere utile per decidere se automatizzare o ricreare il modello con nuovi dati.

Esecuzione dei modelli

In IBM SPSS Modeler l'esecuzione dei modelli è un'attività piuttosto semplice. Dopo aver inserito il nodo del modello nel flusso e modificato i parametri, basta eseguire il modello per produrre risultati visualizzabili. I risultati verranno riportati nel visualizzatore Modelli generati a destra dello spazio di lavoro. È possibile fare clic con il pulsante destro del mouse su un modello per esplorarne i risultati. Per la maggior parte dei modelli, è possibile inserire il modello generato nel flusso per valutare ulteriormente e implementare i risultati. Inoltre, i modelli possono essere salvati in IBM SPSS Modeler per semplificarne il riutilizzo.

Descrizione del modello

Quando si esaminano i risultati di un modello, prendere appunti sulla propria esperienza di modellazione. Questi appunti possono essere archiviati con il modello stesso utilizzando la finestra di dialogo delle annotazioni dei nodi o lo strumento per i progetti.

Elenco delle attività

Per ogni modello, registrare le informazioni seguenti:

- È possibile dedurre conclusioni significative da questo modello?
- Il modello ha rivelato nuove informazioni o schemi inconsueti?
- Si sono verificati problemi di esecuzione per il modello? I tempi di elaborazione sono stati ragionevoli?
- Si sono verificati problemi concernenti la qualità dei dati, ad esempio un alto numero di valori mancanti?
- Sono state verificate incoerenze nei calcoli che dovrebbero essere documentate?

Valutazione del modello

Una volta ottenuti diversi modelli iniziali, esaminarli più attentamente per determinare quali di essi siano sufficientemente accurati ed efficaci da poter essere considerati finali. Il termine "finale" ha diversi significati, ad esempio "pronto per la distribuzione" o "descrittivo di schemi interessanti". Consultare il piano di test precendemente approntato consente di effettuare questa valutazione dal punto di vista dell'azienda.

Valutazione completa del modello

Per ciascun modello considerato, è buona norma eseguire una valutazione sistematica basata sui criteri generati nel piano di test. Questa è la fase in cui è possibile aggiungere il modello generato al flusso e utilizzare grafici di valutazione o nodi di analisi per analizzare l'efficacia dei risultati. È inoltre necessario considerare se i risultati sono sensati o se sono troppo semplicistici per gli obiettivi di business, come, ad esempio, in una sequenza che riveli acquisti del tipo vino > vino > vino.

Una volta eseguita la valutazione, classificare i modelli basandosi su criteri oggettivi (accuratezza del modello) e soggettivi (semplicità di utilizzo o di interpretazione dei risultati).

Elenco delle attività

- Mediante gli strumenti di data mining di IBM SPSS Modeler, quali grafici di valutazione, nodi di analisi e grafici di convalida incrociata, valutare i risultati del modello.
- Condurre una verifica dei risultati basata sulla comprensione del problema di business. Consultare analisti di dati o altri esperti che potrebbero fornire informazioni importanti sulla rilevanza di particolari risultati.
- Valutare se i risultati di un modello sono facilmente implementabili. Chiedersi se l'organizzazione richiede che i risultati siano implementati sul Web o reinviati al data warehouse.
- Analizzare l'impatto dei risultati sui criteri di riuscita. Chiedersi se rispondono agli obiettivi stabiliti durante la fase Business Understanding.

Se tutte le questioni riportate sono state risolte e si ritiene che i modelli correnti soddisfino pienamente gli obiettivi definiti, è tempo di passare a una valutazione più accurata dei modelli e alla distribuzione finale. In caso contrario, avvalersi delle conoscenze acquisite e rieseguire i modelli con impostazioni di parametri rettificate.

Esempio di vendita al dettaglio in linea: valutazione dei modelli

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Miglioramento delle indicazioni. Una delle reti Kohonen e uno dei raggruppamenti a due fasi hanno prodotto entrambi risultati ragionevoli e il rivenditore in linea non riesce a scegliere quale delle due soluzioni adottare. Sul lungo periodo, l'azienda spera di utilizzare entrambe le tecniche, accettando le indicazioni su cui entrambe concordano e studiando più dettagliatamente le situazioni in cui differiscono. Con un piccolo sforzo e applicano le conoscenze di business, il rivenditore può sviluppare ulteriori regole per risolvere le differenze tra le due tecniche.

Il rivenditore ritiene inoltre che i risultati che includono le informazioni sulle sessioni siano sorprendentemente validi. Esistono quindi i presupposti per applicare le indicazioni all'esplorazione del sito. È possibile utilizzare in tempo reale un insieme di regole che definisce i successivi passaggi più probabili dei clienti, per incidere sul contenuto del sito direttamente durante l'esplorazione.

Miglioramento dell'esplorazione del sito. Il modello Sequenza fornisce al rivenditore in linea un livello elevato di confidenza circa la possibilità di prevedere determinati percorsi dei clienti, generando risultati che suggeriscano un numero gestibile di modifiche alla struttura del sito.

Registrazione dei parametri modificati

In base a quanto si è appreso durante la fase di valutazione, è giunto il momento di esaminare nuovamente i modelli. Esistono due opzioni:

- Regolare i parametri dei modelli esistenti.
- Scegliere un modello diverso per risolvere il problema di data mining.

In entrambi i casi, si tornerà all'attività di creazione dei modelli, che verrà ripetuta fino a quando i risultati non saranno soddisfacenti. Ripetere questo passaggio più volte è assolutamente normale. Accade spesso che i data miner valutino e rieseguano i modelli diverse volte prima di scegliere quello che soddisfa le loro esigenze. Questo è un buon motivo per creare più modelli alla volta e confrontarne i risultati prima di regolare i parametri di ognuno di essi.

Considerazioni utili

Prima di passare alla valutazione finale dei modelli, stabilire se la valutazione iniziale è stata sufficientemente accurata.

Elenco delle attività

- Si è in grado di comprendere i risultati dei modelli?
- I risultati dei modelli sono sensati da un punto di vista puramente logico? Esistono apparenti incoerenze che rendono necessaria un'ulteriore esplorazione?

- I risultati sembrano rispondere alla domanda di business?
- Sono stati utilizzati nodi di analisi e grafici guadagno cumulativo e dei profitti per confrontare e valutare l'accuratezza dei modelli?
- Sono stati analizzati diversi tipi di modelli, per confrontarne i risultati?
- I risultati del modello sono implementabili?

Se i risultati della modellazione dei dati sembrano accurati e pertinenti, è giunto il momento di condurre una valutazione più dettagliata prima della distribuzione finale.

Capitolo 6. Valutazione

Panoramica sulla fase di valutazione

A questo punto è stata completata la maggior parte del progetto di data mining. È stato inoltre determinato, nella fase Modeling, che i modelli creati sono tecnicamente corretti ed efficaci, secondo i **criteri per la valutazione dell'esito del data mining** definiti in precedenza.

Prima di continuare, tuttavia, è necessario valutare i risultati del processo utilizzando i **criteri di riuscita di business** stabiliti all'inizio del progetto. Si tratta di un passaggio chiave per garantire che l'azienda possa avvalersi dei risultati ottenuti. Il data mining ha prodotto due tipi di risultati:

- I **modelli** finali selezionati nella fase precedente di CRISP-DM.
- Eventuali conclusioni o deduzioni derivate dai modelli stessi e dal processo di data mining definiti **punti d'arrivo**.

Valutazione dei risultati

In questa fase, si formalizzerà il giudizio circa l'esito dei risultati del progetto, se, cioè abbiano soddisfatto o meno i criteri di business definiti per valutare la riuscita del progetto. Tale passaggio richiede una comprensione chiara degli obiettivi di business dichiarati. Nella valutazione del progetto occorre quindi includere i responsabili delle decisioni chiave.

Elenco delle attività

Innanzitutto, è necessario documentare la valutazione per determinare se i risultati data mining soddisfano i criteri di business. Per la stesura del report, tenere presenti le domande seguenti:

- I risultati sono espressi chiaramente e in una forma facilmente presentabile?
- Si è giunti a conclusioni particolarmente originali e insolite che debbano essere evidenziate?
- È possibile classificare i modelli e le conclusioni in base alla loro applicabilità agli obiettivi di business?
- In generale, in che grado i risultati acquisiti rispondono agli obiettivi di business?
- Quali ulteriori domande hanno fatto emergere i risultati ottenuti? Come è possibile esprimere tali domande in termini di business?

Dopo aver valutato i risultati, compilare un elenco di modelli approvati da includere nel report finale. Tale elenco deve includere i modelli che soddisfano sia gli obiettivi di data mining che quelli di business.

Esempio di vendita al dettaglio in linea: valutazione dei risultati

Scenario di Web-mining in cui viene utilizzato CRISP-DM

I risultati complessivi della prima esperienza di data mining del rivenditore in linea sono discretamente semplici da comunicare da un punto di vista di business: lo studio ha prodotto quelle che si spera siano le migliori indicazioni sui prodotti e ha migliorato la struttura del sito. Le modifiche alla struttura del sito sono basate sulle sequenze di esplorazione dei clienti, che rivelano le pagine che i clienti desiderano visualizzare ma che richiedono diversi passaggi per essere raggiunte. L'evidenza che le indicazioni sui prodotti sono state migliorate è più difficile da comunicare, poiché le regole decisionali possono diventare complicate. Per produrre il report finale, gli analisti tenteranno di identificare alcune tendenze generali negli insiemi di regole che possono essere più facilmente illustrate.

Classificazione dei modelli. Poiché molti dei modelli iniziali sembravano essere significativi da un punto di vista di business, la classificazione all'interno del gruppo è stata basata su criteri statistici, semplicità di interpretazione e diversità. In questo modo, il modello ha fornito diverse indicazioni per una varietà di situazioni.

Nuove domande. La domanda più importante cui lo studio deve fornire una risposta concerne il modo in cui il rivenditore in linea può comprendere meglio i suoi clienti. Le informazioni presenti nel database dei clienti giocano un ruolo determinante per la creazione di gruppi per le indicazioni. Sebbene esistano delle specifiche regole per fornire suggerimenti ai clienti di cui non sono disponibili informazioni, tali suggerimenti hanno un carattere più generale di quelli che possono essere offerti ai clienti registrati.

Rivedi processo

Le metodologie efficaci includono in genere dei tempi dedicati alla riflessione sui punti di forza e di debolezza del processo appena concluso. Il data mining non fa eccezione. Una delle parti di CRISP-DM consiste nell'imparare dall'esperienza fatta, in modo che i futuri progetti di data mining risultino più efficaci.

Elenco delle attività

Innanzitutto, è necessario riepilogare le attività e le decisioni di ogni fase, inclusi i passaggi di preparazione dei dati, la creazione del modello e così via. Quindi, per ogni fase, tenere presenti le seguenti domande e indicare suggerimenti per migliorare la situazione:

- Questa fase ha contribuito al valore dei risultati finali?
- Esiste un modo per ottimizzare o migliorare questa particolare fase o operazione?
- Quali sono stati gli errori di questa fase? Come è possibile evitarli in futuro?
- Ci si è trovati in vicoli ciechi, ad esempio, a causa di modelli che si sono dimostrati infruttuosi? Esiste un modo per prevedere tali situazioni senza sbocchi, al fine di concentrare gli sforzi e renderli più efficaci?
- Si sono verificate sorprese (sia buone che cattive) in questa fase? A posteriori, esiste un modo ovvio per prevedere tali sorprese?
- Esistono decisioni o strategie alternative che avrebbero potuto essere adottate in una determinata fase? Prendere nota di tali alternative per i futuri progetti di data mining.

Esempio di vendita al dettaglio in linea: report di verifica

Scenario di Web-mining in cui viene utilizzato CRISP-DM

La verifica del processo del progetto di data mining iniziale ha consentito al rivenditore in linea di comprendere meglio le interrelazioni tra i passaggi del processo. Inizialmente riluttante a "ritornare sui suoi passi" nel processo di CRISP-DM, il rivenditore riconosce ora che la natura ciclica del processo ne incrementa la forza. La verifica del processo ha inoltre consentito al rivenditore di comprendere quanto segue:

- Quando si verifica un evento inconsueto in una delle altre fasi del processo CRISP-DM, è sempre possibile tornare alla fase di esplorazione.
- La preparazione dei dati, in particolare dei log Web, richiede pazienza, poiché può durare molto a lungo.
- È fondamentale restare concentrati sul problema di business, poiché, una volta preparati i dati per l'analisi, è sin troppo facile iniziare a costruire modelli senza uno sguardo al quadro d'insieme.
- Dopo la fase di modellazione, il processo di Business Understanding ricopre un ruolo ancora più importante per decidere come implementare i risultati e determinare quali ulteriori studi sono garantiti.

Individuazione dei passaggi successivi

Dopo aver prodotto risultati e valutato le esperienze del data mining è lecito chiedersi **quali siano i passaggi successivi**. Questa fase consente di rispondere alla domanda posta alla luce degli obiettivi di business di data mining. A questo punto sono possibili due opzioni:

- **Passare alla fase di distribuzione.** Questa fase consentirà di incorporare i risultati dei modelli nel processo di business e di produrre un report finale. Anche se le operazioni di data mining non hanno

avuto esito positivo, è necessario utilizzare la fase di distribuzione di CRISP-DM per creare un report finale da distribuire allo sponsor del progetto.

- **Tornare indietro e ridefinire o sostituire i modelli.** Se si ritiene che i risultati siano soddisfacenti ma non ottimali, valutare l'ipotesi di rieseguire la modellazione. È possibile avvalersi delle conoscenze acquisite nella fase di valutazione per ridefinire i modelli e produrre risultati migliori.

La decisione deve fondarsi su una valutazione dell'accuratezza e della rilevanza dei risultati della modellazione. Se i risultati rispondono agli obiettivi di business e di data mining, allora si è pronti a passare alla fase di distribuzione. Qualsiasi sia la decisione, documentare con precisione il processo di valutazione.

Esempio di vendita al dettaglio in linea: passaggi successivi

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Il rivenditore in linea è pienamente convinto dell'accuratezza e della rilevanza dei risultati del progetto e decide quindi di passare alla fase di distribuzione.

Al contempo, il team di progetto è pronto a tornare indietro per includere in alcuni dei modelli tecniche di previsione. A questo punto, tutti aspettano la distribuzione dei report finali e il verdetto dei responsabili delle decisioni.

Capitolo 7. Distribuzione

Panoramica sulla fase Distribuzione

La distribuzione è il processo che consente di utilizzare i risultati acquisiti per apportare miglioramenti all'organizzazione. Ciò può implicare un'integrazione formale, come l'implementazione di un modello IBM SPSS Modeler per generare punteggi del tasso di abbandono, che vengono poi letti in un data warehouse. In alternativa, la distribuzione può implicare l'utilizzo delle informazioni ottenute dal data mining per dedurre un cambiamento nell'organizzazione. Se si riscontrano, ad esempio, schemi allarmanti nei dati che indicano un cambiamento di comportamento nei clienti che hanno superato i 30 anni, questi risultati potrebbero non essere formalmente integrati nei sistemi informativi, ma saranno certamente utili per pianificare strategie e prendere decisioni di marketing.

In generale, la fase di distribuzione di CRISP-DM include due tipi di attività:

- Pianificazione e monitoraggio della distribuzione dei risultati
- Completamento di attività riepilogative, quali la produzione di un report finale e la conduzione di una verifica del progetto

A seconda dei requisiti aziendali, potrebbe essere necessario completare uno o entrambi i passaggi.

Pianificazione della distribuzione

Anche se si è ansiosi di condividere i frutti del processo di data mining, dedicare del tempo alla pianificazione di un processo di distribuzione dei risultati ben congegnato e completo.

Elenco delle attività

- Il primo passaggio consiste nel riepilogare i risultati, siano essi modelli o conclusioni. Questo consente di determinare quali modelli possano essere integrati nei sistemi di database e quali conclusioni possano essere presentate ai colleghi.
- Per ciascun modello implementabile, creare un piano dettagliato di distribuzione e integrazione con i sistemi. Prendere nota di tutti i dettagli tecnici, come i requisiti di database per l'output del modello. Può accadere, ad esempio, che il sistema richieda che l'output di modellazione sia implementato in un formato delimitato da tabulazione.
- Per ogni risultato conclusivo, creare un piano di divulgazione delle informazioni agli esperti di strategie.
- Stabilire piani di distribuzione per entrambi i tipi di risultati degni di menzione.
- Valutare la modalità di monitoraggio della distribuzione. Definire, ad esempio, come verrà aggiornato un modello implementato mediante IBM SPSS Modeler Solution Publisher. e quale criterio verrà adoperato per stabilire quando un modello non è più applicabile.
- Identificare eventuali problemi di distribuzione e prevedere gli eventi contingenti. Ad esempio, i responsabili delle decisioni potrebbero volere altre informazioni sui risultati della modellazione e richiedere che siano loro forniti ulteriori dettagli tecnici.

Esempio di vendita al dettaglio in linea: pianificazione della distribuzione

Scenario di Web-mining in cui viene utilizzato CRISP-DM

La corretta distribuzione dei risultati di data mining del rivenditore in linea richiede che le persone giuste ricevano le informazioni appropriate.

Responsabili delle decisioni. I responsabili delle decisioni devono essere informati delle indicazioni e delle modifiche al sito proposte e deve essere loro fornita una breve spiegazione dei benefici che tali

modifiche apporteranno. Ammesso che accettino i risultati dello studio, sarà poi necessario informare le persone che si occuperanno dell'implementazione delle modifiche.

Sviluppatori Web. Coloro che gestiscono il sito Web dovranno incorporare le nuove indicazioni e la nuova organizzazione del contenuto del sito. Gli sviluppatori devono, inoltre, essere informati delle modifiche che *potrebbero* intervenire in futuro in seguito a successivi studi, in modo che possano preparare il terreno per tempo. Consentire al team di prepararsi per la costruzione del sito al volo, basandosi su un'analisi delle sequenze in tempo reale, può risultare utile in seguito.

Esperti di database. Le persone che gestiscono i database dei clienti, degli acquisti e dei prodotti devono essere tenute informate del modo in cui le informazioni dei database verranno utilizzate e di quali attributi potrebbero essere aggiunti ai database nei progetti futuri.

La cosa più importante, comunque, è che il team di progetto sia sempre in contatto con ciascuno di questi gruppi di persone, per coordinare la distribuzione dei risultati e la pianificazione per i progetti futuri.

Pianificazione delle attività di monitoraggio e manutenzione

In un sistema consolidato di distribuzione e integrazione dei risultati della modellazione, il lavoro di data mining può essere svolto in maniera continuativa. Ad esempio, se viene implementato un modello per prevedere le sequenze di acquisti e-basket, è probabile che tale modello debba essere valutato periodicamente, per garantirne l'efficacia e apportare miglioramenti continui. Allo stesso modo, un modello implementato per aumentare la fidelizzazione dei clienti importanti dovrà essere ritoccato quando si sarà raggiunto un determinato livello di fidelizzazione. Il modello potrebbe quindi essere modificato e riutilizzato per fidelizzare clienti a un livello inferiore ma comunque redditizio della piramide dei valori.

Elenco delle attività

Prendere nota dei seguenti problemi e includerli nel report finale.

- Per ogni modello o conclusione, quali fattori o influssi (ad esempio, il valore di mercato o la variazione stagionale) devono essere registrati?
- Come possono essere misurate e monitorate la validità e l'accuratezza di ogni modello?
- Come si potrà stabilire quando un modello è "scaduto"? Fornire, ad esempio, indicazioni specifiche sulle soglie di accuratezza o le modifiche ai dati previste.
- Cosa accadrà alla scadenza di un modello? Sarà possibile semplicemente ricrearlo con nuovi dati o apportare piccole rettifiche? Oppure le modifiche saranno così pervasive da rendere necessario un nuovo progetto di data mining?
- Una volta scaduto, il modello può essere utilizzato per problemi di business simili? È questo il punto in cui una buona documentazione diventa strategica per la valutazione dell'obiettivo di business in relazione a ogni progetto di data mining.

Esempio di vendita al dettaglio in linea: monitoraggio e manutenzione

Scenario di Web-mining in cui viene utilizzato CRISP-DM

La prima attività di monitoraggio consiste nel determinare se la nuova organizzazione del sito e le indicazioni migliorate funzionino realmente. In altre parole, occorre chiedersi se gli utenti siano in grado di arrivare in maniera più diretta alle pagine che stanno cercando e se le vendite incrociate degli articoli consigliati siano aumentate. Dopo alcune settimane di monitoraggio, il rivenditore in linea sarà in grado di determinare la riuscita dello studio.

Ciò che può essere gestito automaticamente è l'inclusione dei nuovi utenti registrati. Quando i clienti si registrano sul sito, gli insiemi di regole correnti possono essere applicati alle informazioni immesse, per determinare quali indicazioni fornire.

Decidere quando aggiornare gli insiemi di regole, per determinare le indicazioni da fornire, è un'attività un po' più complessa. L'aggiornamento degli insiemi di regole non è un processo automatico, poiché la creazione di gruppi richiede l'intervento umano per stabilire l'adeguatezza di una determinata soluzione.

Dal momento che i progetti futuri richiedono modelli più complessi, la necessità e la portata delle operazioni di monitoraggio sono destinate a crescere. Se possibile, la maggior parte del processo di monitoraggio dovrebbe essere automatica con report pianificati regolarmente e sempre esaminabili. In alternativa, l'azienda potrebbe orientarsi verso la creazione di modelli che forniscano previsioni al volo. Questa seconda scelta richiede procedure più sofisticate di quelle del primo progetto di data mining.

Creazione di un report finale

Creare un report finale serve non solo a collegare conclusioni disgiunte nella precedente documentazione, ma anche a comunicare i risultati ottenuti. Sebbene possa sembrare ovvio, è importante presentare i risultati ai diversi soggetti coinvolti nel processo. Tali soggetti possono includere amministratori tecnici, che saranno responsabili dell'implementazione dei risultati di modellazione, e sponsor di marketing e gestione, che baseranno le loro decisioni sui risultati in questione.

Elenco delle attività

Innanzitutto, occorre valutare il ruolo ricoperto dai destinatari del report, se si tratta di sviluppatori tecnici o di responsabili concentrati sul mercato. Potrebbe essere necessario creare report separati per ogni gruppo di destinatari, se le loro esigenze divergono. In ogni caso, il report deve includere la maggior parte dei punti seguenti:

- Una descrizione accurata del problema di business originale
- Il processo utilizzato per condurre il data mining
- Costi del progetto
- Note su qualsiasi deviazione dal piano di progetto originale
- Un riepilogo dei risultati di data mining: modelli e conclusioni
- Una panoramica del piano di distribuzione proposto
- Indicazioni per l'ulteriore lavoro di data mining, inclusi interessanti indizi emersi durante l'esplorazione e la modellazione

Preparazione di una presentazione finale

Oltre al report di progetto, potrebbe essere necessario presentare le conclusioni del progetto a un team di sponsor o reparti correlati. In questo caso, è possibile utilizzare le stesse informazioni del report, presentate però in una prospettiva più ampia. I diagrammi e i grafici di IBM SPSS Modeler possono essere facilmente esportati per questo tipo di presentazione.

Esempio di vendita al dettaglio in linea: report finale

Scenario di Web-mining in cui viene utilizzato CRISP-DM

La deviazione principale dal piano di progetto originale è un fenomeno interessante per l'ulteriore lavoro di data mining. Il piano originale prevedeva che si individuasse una strategia per fare in modo che i clienti trascorressero più tempo sul sito e visualizzassero più pagine per visita.

Tuttavia, soddisfare un cliente non significa semplicemente trattenerlo in linea. Le distribuzioni del tempo trascorso sul sito per sessione, suddiviso in base agli esiti delle sessioni, hanno evidenziato che la durata della maggior parte delle sessioni che hanno prodotto un acquisto si colloca tra le durate di due gruppi di sessioni che non hanno prodotto acquisti.

Una volta rilevato questo dato, il problema è comprendere se i clienti che hanno trascorso molto tempo sul sito senza acquistare alcun prodotto hanno semplicemente esplorato le pagine o non sono stati in grado di trovare ciò che cercavano. Il passaggio successivo consiste nel capire come distribuire ciò che i clienti stanno cercando per incoraggiare gli acquisiti.

Verifica finale del progetto

Questo è l'ultimo passaggio del modello CRISP-DM e offre la possibilità di formulare impressioni conclusive e confrontare le lezioni apprese durante il processo di data mining.

Elenco delle attività

È necessario intervistare brevemente i soggetti significativamente coinvolti nel processo di data mining. Tra le domande da porre durante le interviste sono incluse le seguenti:

- Quali sono le impressioni globali del progetto?
- Cosa si è appreso durante il processo, sia sul data mining in generale che sui dati disponibili?
- Quali parti del progetto andavano bene? Dove sono sorte le difficoltà? Erano disponibili informazioni che avrebbero potuto ridurre la confusione?

Dopo l'implementazione dei risultati del data mining, è possibile intervistare anche coloro su cui tali risultati hanno effetto, ad esempio i clienti o i business partner. L'obiettivo deve essere quello di determinare se il progetto è stato proficuo e se ha comportato i vantaggi che ci era prefissati di offrire.

I risultati di queste interviste possono essere riepilogati, insieme alle proprie impressioni sul progetto, in un report finale, che deve concentrarsi sulle lezioni apprese dall'esperienza di data mining degli archivi.

Esempio di vendita al dettaglio in linea: verifica finale

Scenario di Web-mining in cui viene utilizzato CRISP-DM

Interviste ai partecipanti al progetto. Il rivenditore in linea rileva che i partecipanti al progetto più strettamente associati allo studio dall'inizio alla fine sono in prevalenza entusiasti dei risultati e pensano già a progetti futuri. Il gruppo degli esperti di database sembra cautamente ottimista; sebbene riconosca l'utilità dello studio, ne evidenzia l'ulteriore carico sulle risorse dei database. Durante lo studio era disponibile un consulente, ma andando avanti ed espandendosi l'ambito del progetto, sarà necessario un altro dipendente dedicato alla manutenzione dei database.

Interviste ai clienti. Il feedback dei clienti è stato largamente positivo. Uno dei problemi su cui non si è riflettuto abbastanza è stato l'impatto della modifica alla struttura del sito sui clienti consolidati. In pochi anni, i clienti registrati hanno sviluppato determinate aspettative sulla modalità di organizzazione del sito. Il feedback degli utenti registrati non è positivo quanto quello dei clienti non registrati e alcuni di loro non apprezzano affatto le modifiche. Il rivenditore in linea deve considerare attentamente il problema e valutare se una modifica produrrà un numero tale di nuovi clienti da poter rischiare di perdere quelli esistenti.

Note

Queste informazioni sono state sviluppate per prodotti e servizi offerti negli Stati Uniti. Questo materiale potrebbe essere disponibile da IBM in altre lingue. Tuttavia, potrebbe essere necessario disporre di una propria copia del prodotto o versione di prodotto in quella lingua per potervi accedere.

IBM può non offrire i prodotti, i servizi o le funzioni presentati in questo documento in altri paesi. Consultare il rappresentante locale IBM per le informazioni sui prodotti e servizi attualmente disponibili nella propria zona. Qualsiasi riferimento ad un prodotto, programma o servizio IBM non implica o intende dichiarare che solo quel prodotto, programma o servizio IBM può essere utilizzato. In sostituzione a quelli forniti da IBM, è possibile utilizzare prodotti, programmi o servizi funzionalmente equivalenti che non comportino violazione dei diritti di proprietà intellettuale o di altri diritti IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può avere applicazioni di brevetti o brevetti in corso relativi all'argomento descritto in questo documento. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Chi desiderasse ricevere informazioni relative a licenze può rivolgersi per iscritto a:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Per richieste di licenze relative ad informazioni double-byte (DBCS) contattare il Dipartimento di Proprietà Intellettuale IBM nel proprio paese o inviare richieste per iscritto a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM (INTERNATIONAL BUSINESS MACHINES CORPORATION) FORNISCE LA PRESENTE PUBBLICAZIONE "NELLO STATO IN CUI SI TROVA" SENZA GARANZIE DI ALCUN TIPO, ESPRESSE O IMPLICITE, IVI INCLUSE, A TITOLO DI ESEMPIO, GARANZIE IMPLICITE DI NON VIOLAZIONE, DI COMMERCIABILITÀ E DI IDONEITÀ PER UNO SCOPO PARTICOLARE. Alcune giurisdizioni non escludono le garanzie implicite; di conseguenza la suddetta esclusione potrebbe, in questo caso, non essere applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM si riserva il diritto di apportare miglioramenti e/o modifiche al prodotto o al programma descritto nel manuale in qualsiasi momento e senza preavviso.

Tutti i riferimenti a siti Web non IBM sono forniti unicamente a scopo di consultazione e non devono essere in alcun modo considerati come complementari a tali siti Web. I matteriali disponibili su tali siti Web non fanno parte del materiale relativo a questo prodotto IBM e l'utilizzo di questi è a discrezione dell'utente.

IBM può utilizzare o distribuire qualsiasi informazione fornita dall'utente nel modo che ritiene più idoneo senza incorrere in alcun obbligo nei confronti dell'utente stesso.

Coloro che detengono la licenza su questo programma e desiderano avere informazioni su di esso allo scopo di consentire: (i) lo scambio di informazioni tra programmi indipendenti ed altri (compreso questo) e (ii) l'uso reciproco di tali informazioni dovrebbero contattare:

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Queste informazioni possono essere rese disponibili secondo condizioni contrattuali appropriate, compreso, in alcuni casi, l'addebito di un canone.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale concesso in licenza disponibile sono forniti da IBM in base ai termini dell'IBM Customer Agreement, dell'IBM International Program License Agreement o di qualsiasi altro accordo equivalente tra le parti.

I dati delle prestazioni e gli esempi client citati vengono presentati solo a scopo illustrativo. I risultati delle prestazioni effettive possono variare in base alle configurazioni specifiche e alle condizioni di funzionamento.

Le informazioni relative a prodotti non IBM sono ottenute dai fornitori di quei prodotti, dagli annunci pubblicati o da altre fonti disponibili al pubblico. IBM non ha testato quei prodotti e non può garantire l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Commenti relativi alle prestazioni di prodotti non IBM, dovrebbero essere indirizzati ai fornitori di questi prodotti.

Qualsiasi affermazione relativa agli obiettivi e alla direzione futura di IBM è soggetta a modifica o revoca senza preavviso e concerne esclusivamente gli scopi dell'azienda.

Questa pubblicazione contiene esempi di dati e prospetti utilizzati quotidianamente nelle operazioni aziendali. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e qualsiasi somiglianza a persone o aziende commerciali reali è puramente casuale.

Marchi

IBM, il logo IBM e ibm.com sono marchi o marchi registrati di International Business Machines Corp., registrati in numerose giurisdizioni del mondo. I nomi di altri prodotti e servizi potrebbero essere marchi di IBM o di altre società. Per un elenco aggiornato di marchi IBM, consultare il web nella sezione Copyright and trademark information, all'indirizzo www.ibm.com/legal/copytrade.shtml.

Adobe, il logo Adobe logo, PostScript ed il logo PostScript sono marchi o marchi registrati di Adobe Systems Incorporated negli Stati Uniti e/o in altri paesi.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o relative controllate negli Stati Uniti e altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o in altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o in altri paesi.

UNIX è un marchio registrato di Open Group negli Stati Uniti e/o in altri paesi.

Java e tutti i marchi e i logo relativi a Java sono marchi commerciali o marchi registrati di Oracle e/o delle sue affiliate.

Termini e condizioni per la documentazione del prodotto

Le autorizzazioni per l'uso delle presenti pubblicazioni sono concesse in conformità con i seguenti termini e condizioni.

Applicabilità

I termini e le condizioni riportati di seguito si aggiungono alle condizioni di utilizzo per il sito Web IBM.

Uso personale

È possibile riprodurre tali pubblicazioni per uso personale e non commerciale nel rispetto di tutte le informazioni relative alla proprietà. Non è possibile distribuire, visualizzare o utilizzare tali pubblicazioni, o una parte di esse, senza l'esplicito consenso di IBM.

Uso commerciale

È possibile riprodurre, distribuire e visualizzare queste pubblicazioni unicamente all'interno del proprio gruppo aziendale a condizione che vengano conservate tutte le indicazioni relative alla proprietà. Non è possibile effettuare lavori derivati di queste pubblicazioni o riprodurre, distribuire o visualizzare queste pubblicazioni o qualsiasi loro parte al di fuori del proprio gruppo aziendale senza chiaro consenso da parte di IBM.

Diritti

Fatto salvo quanto espressamente concesso in questa autorizzazione, non sono concesse altre autorizzazioni, licenze o diritti, espressi o impliciti, relativi alle pubblicazioni o a qualsiasi informazione, dato, software o altra proprietà intellettuale qui contenuta.

IBM si riserva il diritto di ritirare le autorizzazioni qui concesse qualora, a propria discrezione, l'utilizzo di queste Pubblicazioni sia a danno dei propri interessi o, come determinato da IBM, qualora non siano rispettate in modo appropriato le suddette istruzioni.

Non è consentito scaricare, esportare o riesportare queste informazioni se non nei limiti stabiliti dalle leggi e normative applicabili, ivi comprese tutte le leggi e le normative sull'esportazione vigenti negli Stati Uniti.

IBM NON GARANTISCE IL CONTENUTO DI QUESTE PUBBLICAZIONI. ESSE SONO FORNITE "NELLO STATO IN CUI SI TROVANO", SENZA GARANZIE DI ALCUN TIPO, ESPRESSE O IMPLICITE, INCLUSE, A TITOLO ESEMPLIFICATIVO, GARANZIE DI COMMERCIABILITA', IDONEITA' PER UNO SCOPO SPECIFICO E DI NON VIOLAZIONE.

Indice analitico

A

accodamento dei dati 22
addestramento/test 26
aggregazione 22
algoritmi 26
analisi dei costi/benefici 9
attributi
    derivazione 21
    selezione 19

B

bontà 26
Business Understanding 5

C

comprensione
    dati 13
    esigenze di business 5
    obiettivi di data mining 9
conclusioni 31
creazione di dati 21
CRISP-DM
    cenni generali 1
    Guida in linea 2
    in IBM SPSS Modeler 2
    risorse aggiuntive 3
criteri
    per la valutazione dell'esito del business 6
    per la valutazione dell'esito del data mining 10
criteri per la valutazione dell'esito positivo
    da un punto di vista di business 6
    dal punto di vista del data mining 9
    in termini tecnici 10

D

data mining
    individuazione dei passaggi successivi 32
    mediante CRISP-DM 1
    verifica del processo 32
Data Preparation 19
Data Understanding 13
dati
    analisi della qualità 16
    attributi 13
    creazione di nuovi dati 21
    descrizione 14
    esclusione 19
    esplorazione 15
    file flat 17
    format 15
    formattazione per la modellazione 23
    integrazione 22
dati *(Continua)*
    ordinamento 23
    partizionamento 26
    pulitura 20
    raccolta 13
    report di raccolta 14
    report sulla qualità 17
    selezione 19
    selezione di attributi 19
    statistiche sulle dimensioni 14
    tipi 13
    unione 22
    valori mancanti 16
    verifica della qualità 16
    visualizzazione 15
definizione
    terminologia del progetto 9
delimitatori 17
dimensione
    insiemi di dati 14
distribuzione 35

E

errori 20
esempi
    fase Business Understanding 5, 7, 10, 11
    fase Data Preparation 19–22
    fase Data Understanding 13–16
    fase di valutazione 31–33
    fase Modeling 25, 27, 29
    vendita al dettaglio in linea 22

F

fase
    Business Understanding 5
    Data Preparation 19
    Data Understanding 13
    modellazione 25
    valutazione 31
file flat 17

G

guida
    CRISP-DM 2

H

HTML
    generazione di report 2

I

ipotesi

ipotesi *(Continua)*
formulazione 16

M

manuali
su CRISP-DM 3
manutenzione 36
metadati 16, 20
modellazione
impostazione delle opzioni 27
preparazione dei dati 19
requisiti dei dati 23
tecniche 25, 26
valutazione dell'output 28
verifica dei risultati 26
modelli
creazione 27
elenco dei modelli approvati 31
non supervisionati 26
parametri 28
supervisionati 26
tipi 28
modelli approvati 31
modelli non supervisionati 26
modelli supervisionati 26
modello
valutazione dei risultati 31
monitoraggio della distribuzione 36

N

nodo Accodamento 22
nodo Crea flag 21
nodo Ricava 21
nodo Unione 22
normalizzazione 21

O

obiettivi
attività coinvolte 6
definizione degli obiettivi di business 5
definizione degli obiettivi di data mining 9
rettifica 16
opzioni
modellazione 28
ordinamento 23
organigrammi 5

P

parametri
modellazione 28, 29
partizionamento 26
pianificazione
distribuzione dei risultati 35
monitoraggio e manutenzione 36
stesura del piano di progetto 10
preparazione dei dati 19
presentazione dei risultati 37
processo
verifica del data mining 32
progetti
conduzione dell'analisi dei costi/benefici 9
creazione del report finale 37
elenco dei rischi e degli imprevisti 8
elenco di requisiti, presupposti e vincoli 8
eseguire una verifica finale 38
inventario delle risorse 7
pulitura dei dati 20
punti d'arrivo 31

Q

qualità
esame dei dati 16
report sulla qualità dei dati 17

R

record
generazione 21
selezione 19
report
descrizione dei dati 15
esplorazione dati 16
generazione dallo strumento per i progetti 2
piano di progetto 10
progetto finale 37
pulitura dei dati 21
qualità dei dati 17
raccolta dei dati 14
requisiti
stesura di un elenco 8
rischi 8
risorse
inventario delle risorse del progetto 7
risorse aggiuntive su CRISP-DM 3
risultati
presentazione 37
valutazione 31
riuscita di business
valutazione dei risultati 31
rumore 17, 20

S

scrittura
piano di progetto 10
report di esplorazione dati 16
report di pulitura dei dati 21
report di raccolta dei dati 14, 15
report sulla qualità dei dati 17
selezione dei dati 19
sfondo
raccolta di informazioni 5
statistiche
analisi 16
statistiche di analisi 16
strumenti
valutazione 10, 11
strumenti di visualizzazione 15
strumento per i progetti 2

T

tecniche
    modellazione 26
terminologia 9
testi descrittivi 2

U

unione dei dati 13, 22

V

valori booleani 14
valori mancanti 13, 16, 20, 21
valori numerici 14
valori simbolici 14
valutazione
    fase di CRISP-DM 31
    individuazione dei passaggi successivi 32
    modelli 28
    situazione di business corrente 7
    strumenti disponibili 10, 11
verifica
    processo di data mining 32
vincoli
    stesura di un elenco 8
vuoti
    raccolta dei dati 13
    verifica della qualità dei dati 16

W

Web-mining
    vendita al dettaglio in linea 5, 7, 10, 19–22, 25, 27, 29, 31–33