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

CRISP-DM 1.0

Metodología CRISP-DM

Este documento describe el proceso de modelado CRISP-DM y contiene la información sobre la metodología de CRISP-DM, el modelo de referencia de CRISP-DM, la guía de usuario de CRISP-DM , Y el reporte CRISP-DM, así como un apéndice con información adicional relacionada. Este documento e información aquí son propiedad exclusiva de los compañeros del consorcio CRISP-DM: NCR Ingeniería de sistemas Copenhague (EE. UU y Dinamarca), DaimlerChrysler AG (Alemania), SPSS Inc. (EE. UU), Y OHRA Verzekeringen en Deposita Groep B.V. (Países Bajos).

Advertencia

El CRISP-DM fue concebido a finales de 1996 por tres "veteranos" del joven e inmaduro mercado de minería de datos. DaimlerChrysler (entonces Daimler-Benz) estaba ya delante de la mayoría de las organizaciones industriales y comerciales en la aplicación de la minería de datos en sus operaciones de negocios.

SPSS (entonces ISL) había estado proporcionando servicios basados en Minería de datos desde 1990 y había lanzado la primer herramienta de trabajo comercial de Minería de Datos Clementine en 1994.

NCR, como parte de su objetivo para entregar valor adicional a su Teradata ® Almacén de datos (data warehouse), habían establecido los equipos consultores de minería de datos y especialistas de tecnología para atender las exigencias de sus clientes.

En aquel tiempo, el temprano interés del mercado en la minería de datos mostraba signos de explosión en la comprensión popular. Esto era tan apasionante como aterrador. Todos nosotros habíamos desarrollado nuestro ingreso (aproximación) a la minería de datos que pasamos de largo. ¿Nosotros hacíamos lo correcto? ¿Cada nueva adopción de minería de datos iba a tener que aprenderse, como nosotros habíamos iniciado, por prueba y error? ¿Y desde la perspectiva de un proveedor, como podíamos manifestarnos a clientes anticipados que la minería de datos era suficientemente madura para ser adoptado como una parte clave de su proceso de negocio?

Un modelo de proceso estándar, pensamos, sin propietarios y libremente disponible, podría dirigir estas cuestiones para nosotros y para todos los profesionales.

Un año más tarde, nosotros habíamos formado un consorcio, inventado una sigla (Proceso Estándar Industrial Híbrido para la Minería de Datos), obtenido financieramente de la Comisión europea, e iniciado para establecer nuestras ideas iniciales. Como el CRISP-DM fue planeado para ser herramienta industrial, y de aplicación neutral, nosotros sabíamos que tuvimos que ser introducidos a una amplia gama como de profesiones y otros (tal como vendedores de almacenes de datos y consultas de administración) con un interés personal en minería de datos. Hicimos esto por crear el Grupo interesado especialmente en CRISP-DM (“el GIS”, como se hizo conocido). Lanzamos el GIS por difusión de una invitación a partes interesadas para unirnos en Amsterdam para todo un día de taller: Nosotros compartiríamos nuestras ideas, los invitaríamos a presentar las suyas, y abrimos la discusión de como tomar el CRISP-DM en adelante.

En los días del taller, había un sentimiento de agitación entre los miembros del consorcio. ¿Alguien estaría bastante interesado en mostrarse? ¿O, si ellos lo hicieran, nos dirían que ellos realmente no vieron una necesidad urgente para un proceso estándar? ¿O que nuestras ideas estaban ahora fuera del paso que cualquier otra idea de estandarización era una fantasía impracticable?

- El taller sobrepasó todas nuestras expectativas. Tres cosas se destacaron:

- El doble de personas apareció de lo que al principio habíamos esperado.

- Había un acuerdo general aplastante que la industria necesitaba de un proceso estándar y lo necesitaba ahora.

- Como los asistentes presentaron sus opiniones sobre la minería de datos desde su experiencia de proyecto, se hizo claro que aunque hubiera diferencias superficiales - principalmente en la demarcación de fases y en la terminología- hubo enormes puntos en común en como ellos vieron el proceso de minería de datos.

Hacia el final del taller, sentimos confidencias que nosotros podríamos entregar, con la entrada del GIs'S y las críticas, un modelo de proceso estándar para atender la comunidad de minería de datos.

Durante los dos próximos años y medio, trabajamos para desarrollar y refinar el CRISP-DM. Controlamos pruebas en vivo, en proyectos de gran escala de minería de datos, en Mercedes Benz y en nuestro compañero del sector de seguros, OHRA. Trabajamos sobre la integración del CRISP-DM con herramientas comerciales de minería de datos. El GIS demostró ser valioso, creciendo a más de 200 miembros y sosteniendo talleres en Londres, Nueva York, y Bruselas.

Hacia el final del proyecto de la parte financiada por CE -mid-1999- habíamos producido lo que consideramos un esbozo de buena calidad del modelo de proceso. Aquellos familiarizados con aquel esbozo encontrarán que un año más tarde, aunque ahora mucho más completo y mejor presentado, el CRISP-DM 1.0 no es en ningún caso radicalmente diferente. Nosotros éramos sumamente conscientes que, durante el proyecto, el modelo de proceso está todavía con muchísimo trabajo-en-progreso; el CRISP-DM sólo había sido validado sobre un juego estrecho de proyectos. A lo largo del año pasado, DaimlerChrysler tenía la oportunidad de aplicar el CRISP-DM a una más amplia gama de usos. Los grupos de Servicios Profesionales del SPSS' Y NCR'S han adoptado el CRISP-DM y usado satisfactoriamente sobre numerosos contratos de cliente cubriendo muchas industrias y problemas de negocio.

En todo este tiempo, hemos visto que los proveedores de servicio de fuera del consorcio adoptan el CRISP-DM, repetidas referencias por los analistas como el estándar real para la industria, y una conciencia creciente de su importancia entre clientes (CRISP-DM esta ahora con frecuencia referido en invitaciones al concurso y en documentos RFP). Creemos que nuestra iniciativa ha sido a fondo reivindicado, y mientras futuras mejoras y extensiones son muy deseables como inevitables, consideramos la versión de CRISP-DM 1.0 suficientemente validado para ser publicado y distribuido.

El CRISP-DM no ha sido construido a una manera teórica, académica que trabaja de principios técnicos, ni hizo comités de la elite de gurúes creando detrás de puertas cerradas. Ambos de estos accesos a metodologías que se desarrollan han sido intentados en el pasado, pero raras veces conducían a lo práctico, lo acertado, y extensamente ha adoptado normas. El CRISP-DM tiene éxito porque esta profundamente basado en la experiencia práctica, la experiencia del mundo real de como la gente conduce proyectos de minería de datos. Y en este sentido, somos abrumadoramente el deudor a muchos médicos quien contribuyeron con sus esfuerzos y sus ideas en todas partes del proyecto.

El consorcio de CRISP-DM

Agosto de 2000

I-Introducción

1. La metodología CRISP-DM

1.1. Interrupción jerárquica

La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos. (Ver la figura 1.)

En el *nivel superior*, el proceso de minería de datos es organizado en un número de fases; cada fase consiste de varias tareas genéricas de segundo nivel. Este *segundo nivel* lo llaman *genérico* porque esta destinado a ser bastante general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible. *Completo* significa que cubre tanto al proceso entero de minería de datos y todas las aplicaciones de minería de datos posibles. *Estable* significa que el modelo debería ser válido para acontecimientos normales y aún para desarrollos imprevistos como técnicas de modelado nuevo.

El *tercer nivel*, el nivel de tarea especializado, es el lugar para describir como las acciones en las tareas genéricas deberían ser realizadas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe como esta tarea se diferencia en situaciones diferentes, como la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo.

La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos.

En la práctica, muchas de las tareas pueden ser realizadas en una orden diferente, y esto a menudo será necesario volver a hacer tareas anteriores repetidamente y repetir ciertas acciones. Nuestro modelo de proceso no intenta capturar todas estas posibles rutas del proceso de la minería de datos porque esto requeriría un modelo de proceso demasiado complejo.

El *cuarto nivel*, la instancia de proceso, es un registro de las acciones, decisiones, y de los resultados de una minería de datos real contratada.

Una instancia de proceso esta organizado según las tareas definidas en los niveles más altos, pero representa lo que en realidad pasó en un contrato particular más bien que lo que pasa en general.

Figura 1: Cuatro niveles de interrupción de la metodología CRISP-DM

1.2. Modelo de referencia y guía de usuario

Horizontalmente, la metodología de CRISP-DM se distingue entre el modelo de referencia y la guía de usuario. El modelo de referencia presenta una descripción rápida de fases, las tareas, y sus salidas, y describen que hacer en el proyecto de minería de datos. La guía de usuario da consejos más detallados e insinuaciones para cada fase y cada tarea dentro de una fase, y representa como realizar un proyecto de minería de datos

Este documento cubre tanto el modelo de referencia como la guía de usuario en el nivel genérico.

2. Pasaje de modelos genéricos a modelos especializados

2.1. Contexto de la minería de datos

El contexto de minería de datos traza un mapa entre lo genérico y el nivel especializado en CRISP-DM. Actualmente, distinguimos entre cuatro dimensiones diferentes de contextos de minería de datos:

- el **dominio de aplicación** es el área específica en la que el proyecto de minería de datos toma lugar
- los **tipos de problemas de minería de datos** describen la(s) clase(s) específica(s) de objetivo(s) con el que el proyecto de minería de datos trata (ver también el Apéndice 2)
- el **aspecto técnico** cubre cuestiones específicas en minería de datos que describe diferentes (técnicas) dificultades que por lo general ocurren durante la minería de datos
- la **herramienta** y las especificaciones de dimensión **técnica** en la que las herramienta(s) de minería de datos y/o técnicas son aplicadas durante el proyecto de minería de datos

La Tabla 1 de abajo resume estas dimensiones de contextos de minería de datos y muestra ejemplos específicos para cada dimensión.

	Data Mining Context			
Dimension	Application Domain	Data Mining Problem Type	Technical Aspect	Tool and Technique
Examples	Response Modeling	Description and Summarization	Missing Values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
	...	Concept Description	...	Decision Tree
		Classification		...
		Prediction		
		Dependency Analysis		

Table 1: Dimensions of data mining contexts and examples

	Contexto de Minería de Datos			
Dimensión	**Dominio de aplicación**	**Tipo de problema de Minería de Datos**	**Aspectos Técnicos**	**Herramienta y Técnica**
Ejemplos	Modelado de respuesta	Descripción y resumen	Valores encontrados	Clementine
	Realimentar predicciones	Segmentación	Salidas	MineSet
	...	Descripción de conceptos	...	Arbol de decisión
		Predicción		...
		Análisis de dependencia		

Tabla 1. Dimensión de contextos y ejemplos de minería de datos

Un contexto específico de minería de datos es un valor concreto para una o más de estas dimensiones. Por ejemplo, un proyecto de minería de datos tratando con un problema de clasificación que se revuelve con la predicción constituye un contexto específico. Lo más específico (los valores) para las dimensiones de contextos diferentes son fijadas (especificadas), lo mas concreto es el contexto de minería de datos.

2.2. Pasaje con contextos

Distinguimos entre dos tipos diferentes de pasajes (plan) entre el nivel genérico y un especializado en el CRISP-DM.

Pasaje para el presente: Si sólo aplicamos el modelo de proceso genérico para realizar un proyecto de minería simple, e intentar pasar de tareas genéricas y sus descripciones al proyecto específico como requerido, hablamos sobre un pasaje solo para (probablemente) un solo uso.

Pasaje para el futuro: Si sistemáticamente especializamos el modelo de proceso genérico según un contexto predefinido (o analizando sistemáticamente de modo similar y consolidando las experiencias de un único proyecto hacia un modelo de proceso especializado para el uso futuro en contextos comparables), hablamos explícitamente de la sobre escritura de un modelo de proceso especializado en términos de CRISP-DM.

Cualquiera de los tipos de trazados es apropiado según sus propios objetivos, depende de su contexto de minería de datos específicos y las necesidades de su organización.

2.3. Pasaje

La estrategia básica para pasar un mapa del modelo de proceso genérico al nivel especializado es la misma para ambos tipos de pasaje:

- Analizar su contexto específico
- Quitar cualquier detalle no aplicable a su contexto

- Agregar cualquier detalle específico a su contexto
- Especializar (o instanciar) el contenido genérico según las características concretas de su contexto
- Renombrar el contenido genérico posible para proporcionar significados más explícitos en su contexto para la aclaración.

3. Descripción de partes

3.1. Contenido

El modelo de proceso de CRISP-DM (este documento) es organizado en cinco partes diferentes:

- **Parte I**: es esta una introducción a la metodología de CRISP-DM, que proporciona algunas directrices generales para pasar un modelo de proceso genérico a modelos de proceso especializados
- **Parte II**: describe el modelo de referencia de CRISP-DM, sus fases, tareas genéricas, y salidas
- **Parte III** presenta la guía de usuario de CRISP-DM, que va más allá de la descripción pura de fases, tareas genéricas, y salidas, y contiene el asesoramiento más detallado sobre como realizar proyectos de minería de datos
- **Parte IV**: Se centra en los informes para ser producidos durante y después de un proyecto, y sugiere contornos para estos informes. Ello también muestra referencias cruzadas entre salidas y tareas.
- **Parte V** es el apéndice, que incluye un glosario de terminología importante y una caracterización de los tipos de problemas de minería de datos

3.2. Objetivo

Los usuarios y los lectores de este documento deberían ser conscientes de las instrucciones siguientes:

- Si usted lee el modelo de proceso de CRISP-DM por primera vez, comience con la Parte I, la introducción, para entender la metodología de CRISP-DM, todos sus conceptos, y como los distintos conceptos se relacionan uno con el otro. En remotas lecturas, usted podría saltar la introducción y sólo verlo si lo necesita para una aclaración.
- Si usted necesita rápido el acceso a una descripción del modelo de proceso de CRISP-DM, referirse a la Parte II, el modelo de referencia de CRISP-DM, otra forma de un proyecto de minería de datos rápidamente o conseguir una introducción a la guía de usuario de CRISP-DM.
- Si usted necesita el asesoramiento detallado en la realización de su proyecto de minería de datos, ver Parte IIII, la guía de usuario de CRISP-DM, es lo más parte más importante de este documento. Nota: si usted no ha leído primero la introducción o el modelo de referencia, vuelva y lea estas primeras dos Partes.
- Si usted está en la etapa de minería de datos cuando usted sobrescribe sus informes, ver Parte IV. Si usted prefiere generar deliberadamente las descripciones durante el proyecto, muévase hacia adelante y hacia atrás entre Partes III y IV como lo desee.
- Finalmente, el apéndice es útil como información adicional de fondo a la MINERÍA de datos y al CRISP-DM. Use el apéndice para buscar varios términos si usted no es aún un experto en el campo.

II-El modelo de referencia CRISP-DM

El modelo de proceso corriente para la minería de datos proporciona una descripción del ciclo de vida del proyecto de minería de datos. Este contiene las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones. Las relaciones podrían existir entre cualquier tarea de minería de datos según los objetivos, el contexto, y –lo más importante- el interés del usuario sobre los datos.

El ciclo de vida del proyecto de minería de datos consiste en seis fases, mostrado en la Figura 2. La secuencia de las fases no es rígida.

El movimiento hacia adelante y hacia atrás entre fases diferentes es siempre requerido. El resultado de cada fase determina que la fase, o la tarea particular de una fase, tienen que ser realizados después. Las flechas indican las más importantes y frecuentes dependencias entre fases.

El círculo externo en la Figura 2 simboliza la naturaleza cíclica de la minería de datos. La minería de datos no se termina una vez que la solución es desplegada. Las informaciones ocultas (lecciones cultas) durante el proceso y la solución desplegada pueden provocar nuevas, a menudo más - preguntas enfocadas en el negocio. Los procesos de minería subsecuentes se beneficiarán de las experiencias previas. En el siguiente, brevemente perfilamos cada fase:

Figura 2: Fases del modelo de referencia CRISP-DM

Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Preparación de datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datos que serán provistos en las herramientas de modelado] de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Modelado

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

Evaluación

En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos.

Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida.

Desarrollo

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

La figura 3 presenta un contexto de fases acompañadas por tareas genéricas y las salidas. En las secciones siguientes, describimos cada tarea genérica y sus salidas más detalladamente. Enfocamos nuestra atención en descripciones de tarea y los resúmenes de salidas.

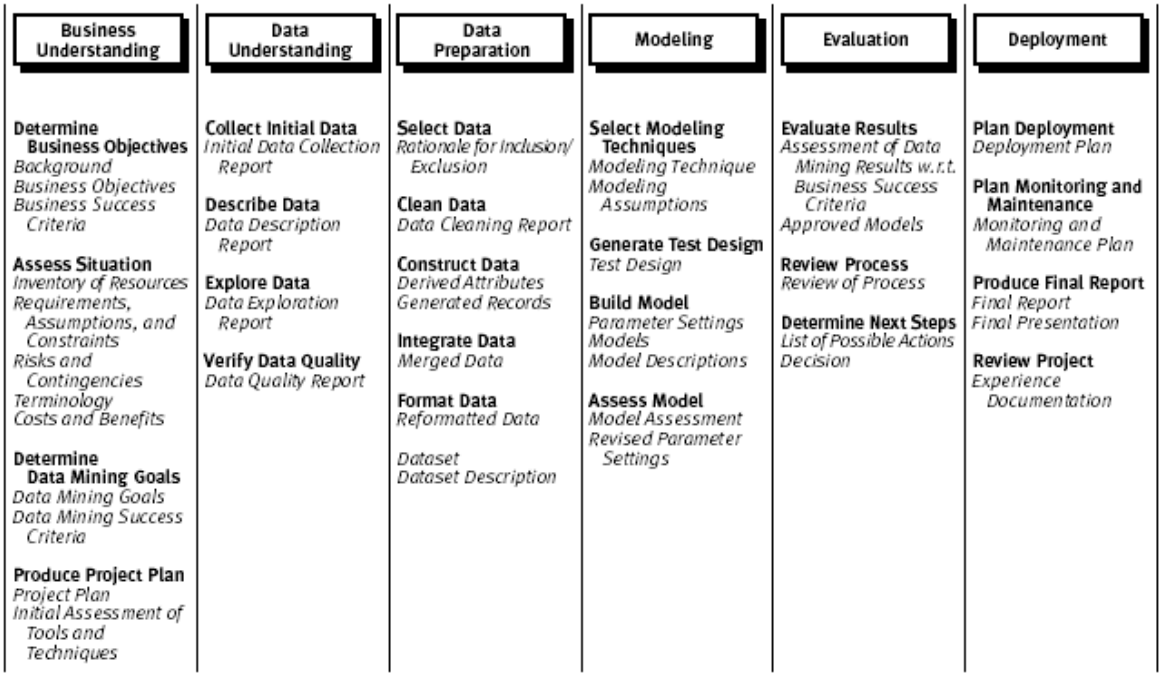

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Figura 3: Tareas genéricas (negritas) y salidas (cursivas) del modelo de referencia CRISP-DM

1. Comprensión del negocio

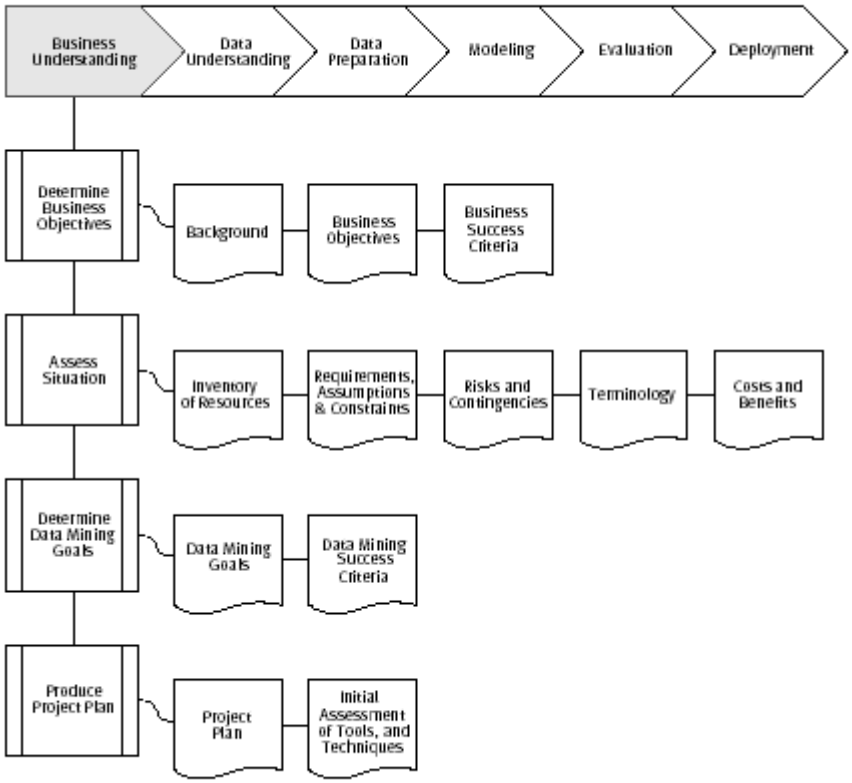

Figure 4: Business Understanding

Figura 4. Comprensión del negocio

1.1. Determinación de objetivos de negocio

Tarea **Determinar los objetivos de negocio**

El primer objetivo del analista de datos para un contexto es entender, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo el cliente tiene muchos objetivos que compiten y restricciones que deben ser correctamente equilibrados. El objetivo del analista debe mostrar (destapar) factores importantes, en el principio, esto puede influir en el resultado del proyecto. Una consecuencia posible de la negligencia de este paso es gastar un gran esfuerzo produciendo respuestas correctas a preguntas incorrectas o erradas.

Salida **Contexto**

Registre la información que conoce sobre la situación de negocio de la organización en el principio del proyecto.

Objetivos de negocio

Describa el objetivo primario del cliente, desde una perspectiva de negocio. Además de los objetivos del negocio primario, allí hay típicamente otras preguntas de negocio relacionadas con lo que al cliente le gustaría administrar. Por ejemplo, el objetivo primario de negocio podría ser mantener a clientes corrientes por predicción cuando ellos son propensos a moverse a un competidor. Los ejemplos de preguntas relacionadas de negocio son "¿Cómo el uso del canal primario (Por ejemplo, ATM, visita al negocio, Internet) afecta si los clientes se quedan o se van? " o "¿Bajar los honorarios de ATM considerablemente reducirá el número de los clientes mas importante que se irán?"

Criterios de éxito de negocio

Describa los criterios para un resultado acertado o útil al proyecto desde el punto de vista del negocio. Esto podría ser bastante específico y capaz de ser medido objetivamente, por ejemplo, la reducción de clientes se revuelve a un cierto nivel o valor, o esto podría ser general y subjetivo, como "dar ideas útiles en las relaciones". En este último caso, esto debería indicarse quien hace el juicio subjetivo.

1.2. Evaluación de la situación

Tarea **Evaluar la situación**

Esta tarea implica la investigación más detallada sobre todos los recursos, restricciones, presunciones, y otros factores que deberían ser considerados en la determinación del objetivo de análisis de datos y el plan de proyecto. En la tarea anterior, su objetivo es para ponerse rápidamente al quid de la situación. Aquí, usted quiere ampliarse sobre los detalles.

Salida **Inventario de recursos**

Listar los recursos disponibles para el proyecto, incluyendo el personal (expertos de negocio, expertos de datos, soportes técnicos, expertos en minería de datos), datos (extractos fijos, aproximaciones a la vida, almacenes de datos, u datos operacionales), recursos computacionales (plataformas de hardware), y software (herramientas de minería de datos, otros software relevantes).

Requerimientos, presunciones, y restricciones

Listar todos los requerimientos del proyecto, incluyendo el programa de terminación, la comprensibilidad y calidad de los resultados, y la seguridad, así como las cuestiones legales. Como parte de esta salida, asegúrese que le permitan usar los datos.

Listar las presunciones hechas por el proyecto. Estas pueden ser presunciones sobre los datos que pueden ser verificados durante la minería de datos, pero también puede incluir presunciones no-comprobables sobre el negocio relacionado con el proyecto. Es en particular importante listar si esto afectará la validez de los resultados.

Listar las restricciones sobre el proyecto. Estas pueden ser restricciones sobre la disponibilidad de recursos, pero puede también incluir coacciones tecnológicas como el tamaño de conjunto de datos lo que es práctico para usar el modelado.

Riesgos y contingencias

Listar los riesgos o los acontecimientos que podrían retrasar el proyecto o hacer que ello falle. Listar los planes de contingencia correspondientes, que acción será tomada si estos riesgos o acontecimientos ocurren.

Terminología

Compile un glosario de terminología relevante al proyecto. Esto puede incluir dos componentes:

(1) Un glosario de terminología relevante del negocio, que forma la parte de la comprensión del negocio disponible al proyecto. La construcción de este glosario es una útil "evocación al conocimiento" y un ejercicio de educación.

(2) Un glosario de terminología de minería de datos, ilustrada con ejemplos relevantes al problema del negocio en cuestión.

Costos y beneficios

Construya un análisis de costo-beneficio para el proyecto, que compare los gastos del proyecto con los beneficios potenciales al negocio si esto es exitoso. La comparación debería ser tan específica como posible. Por ejemplo, use medidas monetarias en una situación comercial.

1.3. Determinación de los objetivos de la minería de datos

Tarea **Determinar los objetivos de la minería de datos**

Un objetivo de negocio declara objetivos en la terminología de negocio. Un objetivo de minería de datos declara objetivos de proyecto en términos técnicos. Por ejemplo, el objetivo de negocio podría ser “Aumentar catálogos de ventas a clientes existentes.” Un objetivo de minería de datos podrían ser “Predecir cuantas baratijas un cliente comprará, obteniendo datos de sus compras de tres años pasados, información demográfica (edad, sueldo, ciudad, etc.), y el precio del artículo.”

Salida **Objetivos de la minería de datos**

Describir las salidas intencionadas del proyecto que permiten el logro de los objetivos de negocio.

Criterios de éxito de la minería de datos

Definir los criterios de un resultado exitoso para el proyecto en términos técnicos -por ejemplo, un cierto nivel de predicción precisa o un perfil de inclinación-a-comprar con un determinado grado de "elevación". Como con un criterio de éxito de negocio, puede ser necesario describir estos en términos subjetivos, en este caso la persona o las personas que hacen el juicio subjetivo deberían ser identificadas.

1.4. Producir el plan del proyecto

Tarea **Producir el plan del proyecto**

Describir el plan intencionado para alcanzar los objetivos de minería de datos y así alcanzar los objetivos de negocio.

El plan debería especificar los pasos para ser realizados durante el resto del proyecto, incluyendo la selección inicial de herramientas y técnicas.

Salida **Plan del Proyecto**

Listar las etapas a ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias. Donde sea posible, haga explícito las iteraciones en gran escala en el proceso de minería de datos -por ejemplo, las repeticiones del modelado y las fases de evaluación.

Como parte del plan de proyecto, es también importante analizar dependencias entre la planificación de tiempo y los riesgos.

Marcar los resultados de estos análisis explícitamente en el plan de proyecto, idealmente con acciones y recomendaciones si los riesgos se manifiestan.

Nota: el plan de proyecto contiene proyectos detallados para cada fase. Decida en este punto que estrategia de evaluación será usada en la fase de evaluación.

El plan de proyecto es un documento dinámico en el sentido de que en el final de cada fase, son necesarios una revisión del progreso y logros y una actualización correspondiente del plan de proyecto es recomendado. Los puntos de revisión específicas para estas actualizaciones son parte del plan de proyecto.

Evaluación inicial de herramientas y técnicas

En la final de la primera fase, una evaluación inicial de herramientas y técnicas debería ser realizada. Aquí, por ejemplo, usted selecciona una herramienta de minería de datos que soporte varios métodos para las distintas etapas del proceso.

Es importante evaluar herramientas y técnicas temprano en el proceso desde la selección de herramientas y técnicas y esto puede influir en el proyecto entero.

2. Comprensión de datos

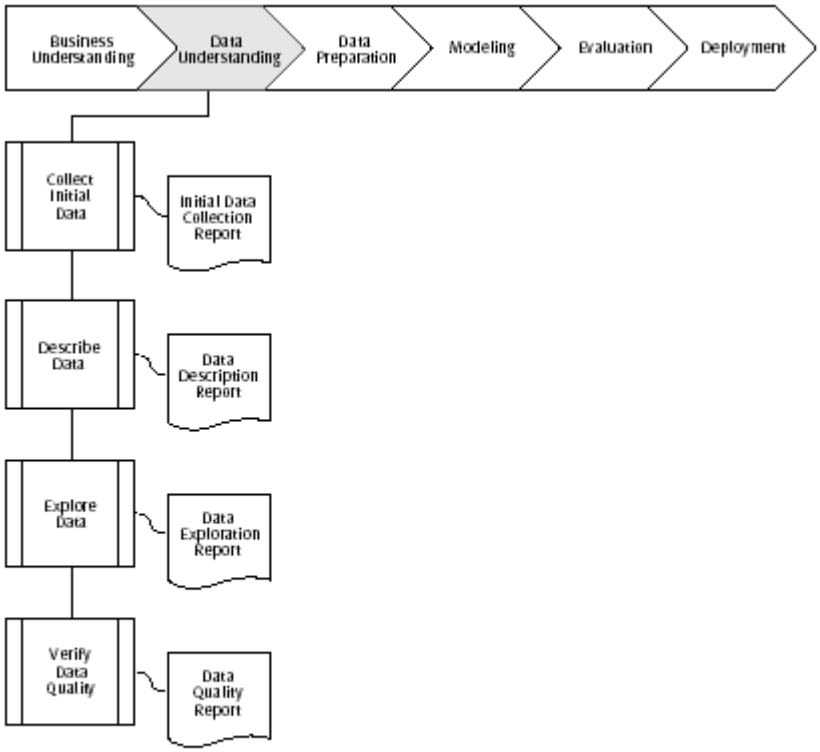

Figure 5: Data understanding

Figura 5: Comprensión de datos

2.1. Recolección de datos iniciales

Tarea **Recolectar datos iniciales**

Adquiera en el proyecto los datos (o el acceso a los datos) listados en los recursos del proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de los datos. Por ejemplo, si usted usa un instrumento específico para la comprensión de los datos, esto perfectamente se entiende para abrir sus datos en esta herramienta. Este esfuerzo posiblemente conduce a los pasos iniciales de preparación de datos.

Note: si usted adquiere datos de múltiples fuentes, la integración es una cuestión adicional, aquí o mas tarde en las fases de preparación de datos más.

Salida **Informe de colección de datos inicial**

Liste el conjunto de dato(s) adquirido(s), juntos con sus posiciones, los métodos usados para adquirirlos, y algunos de los problemas encontrados. Registre los problemas encontrados y algunas de las resoluciones alcanzadas. Esto ayudará con la réplica (observación) futura de este proyecto o con la ejecución de proyectos similares futuros.

2.2. Describir los datos

Tarea **Describir los datos**

Examine las propiedades "gruesas" o "superficiales" de los datos e informe adquiridos en los resultados.

Salida **Informe de descripción de datos**

Describa los datos que han sido adquiridos, incluyendo el formato de los datos, la cantidad de datos (por ejemplo, el número de registros y campos en cada tabla), los identificadores de los campos, y cualquier otro rasgo superficial que ha sido descubierto. Evalúe si los datos adquiridos satisfacen las exigencias relevantes.

2.3. Explorar los datos

Tarea **Explorar los datos**

Esta tarea dirige interrogantes de minería de datos usando preguntas, visualización, y técnicas de reporte. Estos incluyen la distribución de atributos claves (por ejemplo, el atributo objetivo de una tarea de predicción) relacionados entre pares o pequeños números de atributos, los resultados de simples agregaciones, las propiedades de las subpoblaciones significativas, y análisis estadísticos simples. Estos análisis directamente pueden dirigir los objetivos de minería de datos; ellos también pueden contribuir o refinar la descripción de datos e informes de calidad, y alimentar en la transformación y otros pasos de preparación de datos necesarios para análisis futuros.

Salida **Informe de exploración de datos**

Describa los resultados de esta tarea, incluyendo primeras conclusiones o hipótesis iniciales y su impacto sobre el resto del proyecto. Si es apropiado, incluya gráficos y plots para indicar las características de datos que sugieren más examen de subconjuntos de datos interesantes.

2.4. Verificar la calidad de los datos

Tarea **Verificar la calidad de los datos**

Examine la calidad de los datos, dirigiendo preguntas como: ¿Los datos están completos? (¿Esto cubre todo los casos requeridos)? ¿Son correctos, o estos contienen errores y, si hay errores, que tan comunes son estos? ¿Hay valores omitidos en los datos? Si es así, ¿como se representan estos, donde ocurre esto, y que tan comunes son estos?

Salida **Informe de calidad de datos**

Liste los resultados de la verificación de calidad de datos; si existen problemas de calidad, liste las posibles soluciones. Las soluciones a los problemas de calidad de datos generalmente dependen tanto del conocimiento de los datos y como del negocio.

3. Preparación de datos

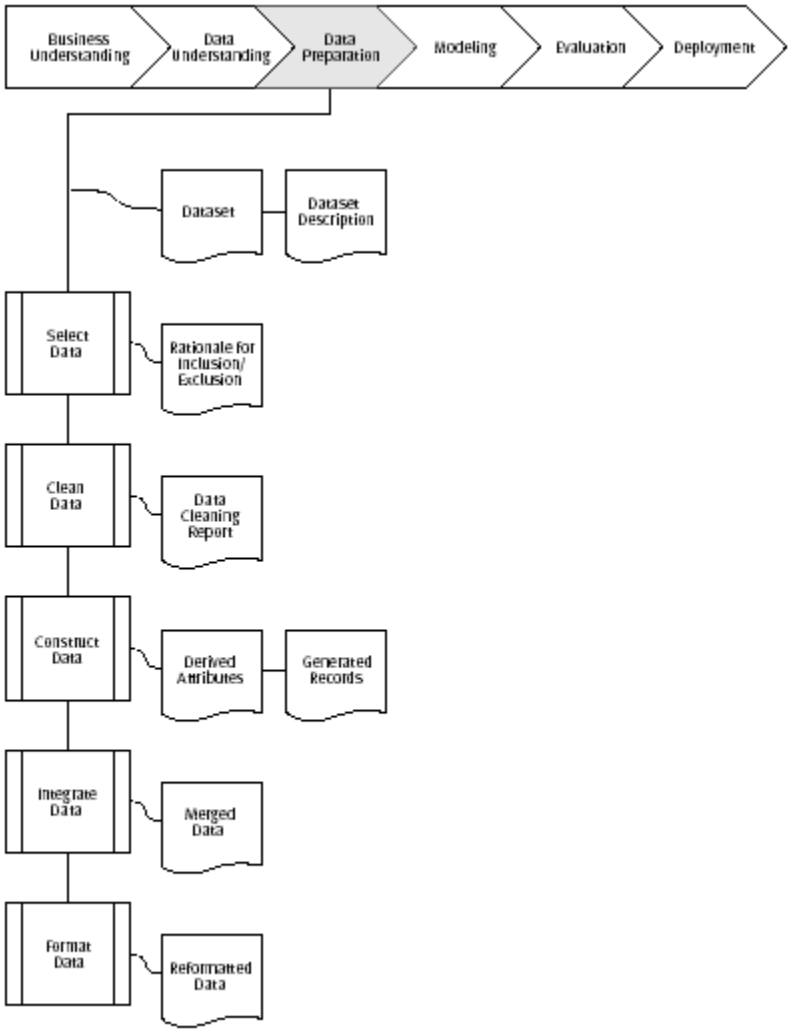

Figure 6: Data preparation

Figura 6: Preparación de datos

Salida **Conjunto de datos**

Este es el conjunto (o conjuntos) producido por la fase de preparación de datos, que será usada para modelar o para el trabajo principal de análisis del proyecto.

Descripción del conjunto de datos

Describir el conjunto de dato (o conjuntos) que será usado para el modelado y el trabajo principal de análisis del proyecto.

3.1. Selección de datos

Tarea **Selección de datos**

Decidir que datos serán usados para el análisis. Los criterios incluyen la importancia a los objetivos de la minería de datos, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Note que la selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla.

Salida **Razonamiento para la inclusión/exclusión**

Listar los datos para ser incluidos/excluidos y los motivos para estas decisiones.

3.2. Limpieza de datos

Tarea **Limpiar datos**

Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de los subconjuntos de datos limpios, la inserción de datos por defectos

adecuados, o técnicas más ambiciosas tales como la estimación de datos faltantes mediante modelado.

Salida **Informe de la limpieza de los datos**

Describa que decisiones y acciones fueron tomadas para dirigir los problemas de calidad de datos informados durante la tarea de Verificación de Calidad de Datos de los Datos de la fase de Comprensión de Datos. Las transformaciones de los datos para una apropiada limpieza y el posible impacto en el análisis de resultados deberían ser considerados.

3.3. Construir datos

Tarea **Construir datos**

Esta tarea incluye la construcción de operaciones de preparación de datos tales como la producción de atributos derivados o el ingreso de nuevos registros, o la transformación de valores para atributos existentes.

Salidas **Atributos derivados**

Los atributos derivados son los atributos nuevos que son construidos de uno o más atributos existentes en el mismo registro. Ejemplo: área = longitud * anchura.

Registros generados

Describa la creación de registros completamente nuevos. Ejemplo: Crear registros para los clientes quienes no hicieron compras durante el año pasado. No había ninguna razón de tener tales registros en los datos brutos, pero para el objetivo del modelado esto podría tener sentido para representar explícitamente el hecho que ciertos clientes no hayan hecho compra nada.

3.4. Integrar datos

Tarea **Integrar datos**

Estos son los métodos por el cual la información es combinada de múltiples tablas o registros para crear nuevos registros o valores.

Salida **Combinación de datos**

La combinación de tablas se refiere a la unión simultánea de dos o más tablas que tienen información diferente sobre el mismo objeto. Ejemplo: una cadena de venta al público tiene una tabla con la información sobre las características generales de cada tienda (Por ejemplo, el espacio, el tipo de comercio), otra tabla con datos resumidos de las ventas (por ejemplo, el beneficio, el cambio porcentual en ventas desde el año anterior), y el otro con información sobre los datos demográficos del área circundante. Cada una de estas tablas contiene un registro para cada tienda. Estas tablas pueden ser combinadas simultáneamente en una nueva tabla con un registro para cada tienda, combinando campos de las tablas fuentes.

Los datos combinados también cubren agregaciones. La agregación se refiere a operaciones en la que nuevos valores son calculados de información resumida de múltiples registros y/o tablas. Por ejemplo, convirtiendo una tabla de compra de clientes donde hay un registro para cada compra en una tabla nueva donde hay un registro para cada cliente, con campos tales como el número de compras, el promedio de la cantidad de compra, el porcentaje de ordenes cobrados a tarjeta de crédito, el porcentaje de artículos bajo promoción, etc.

3.5. Formatear datos

Tarea **Formatear datos**

Formateando transformaciones se refiere a modificaciones principalmente *sintácticas* hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.

Salida **Datos reformateados**

Algunas herramientas tienen requerimientos sobre el orden de los atributos, tales como el primer campo que es un único identificador para cada registro o el último campo es el campo resultado que el modelo debe predecir.

Podría ser importante cambiar el orden de los registros en el conjunto de datos. Quizás la herramienta de modelado requiere que los registros sean clasificados según el valor del atributo de resultado. Comúnmente, los registros del conjunto de datos son ordenados al principio de algún modo, pero el algoritmo que modela necesita que ellos estén en un orden moderadamente arbitrario. Por ejemplo, cuando se usa redes neuronales, esto es generalmente mejor para los registros para ser presentados en un orden aleatorio, aunque algunas herramientas manejen esto automáticamente sin la intervención explicita del usuario.

Además, hay cambios puramente sintácticos hechos para satisfacer las exigencias de la herramienta de modelado específica. Ejemplos: el quitar de comas de adentro de campos de texto en ficheros de datos delimitados por coma, corta todos los valores a un máximo de 32 caracteres.

4. Modelado

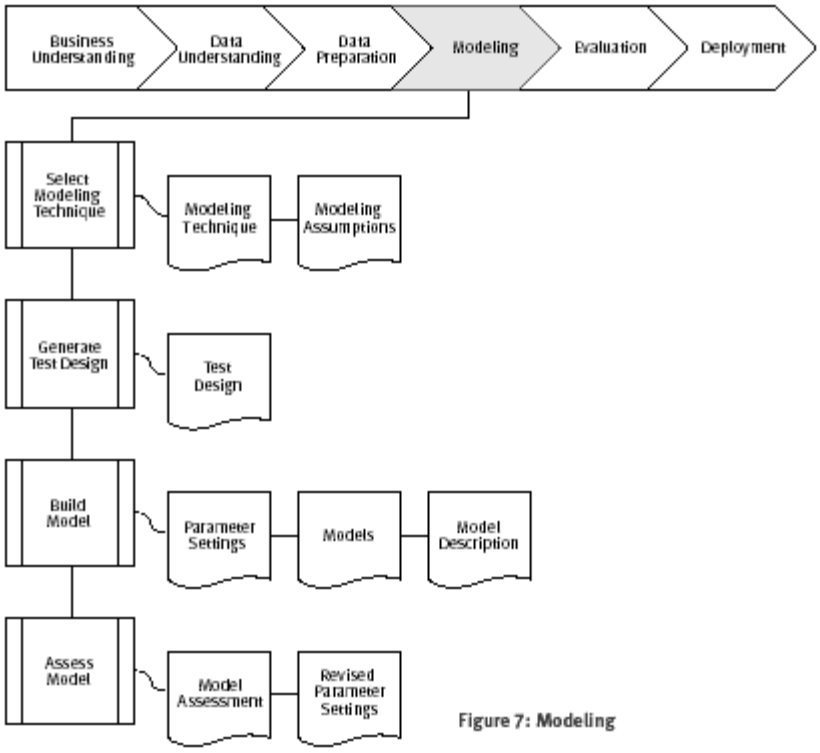

Figura 7: Modelado

4.1. Selección de la técnica de modelado

Tarea **Escoger la técnica de modelado**

Como primer paso en modelado, seleccionar la técnica de modelado real que está por ser usado. Aunque usted haya podido seleccionar una herramienta durante la fase de Comprensión del negocio, esta tarea se refiere a la técnica de modelado específico, por ejemplo, un árbol decisión construido con C4.5, o la generación de red neuronales Back-Propagación. Si múltiples técnicas son aplicadas, se realizan esta tarea separadamente para cada técnica.

Salida **Técnicas de modelado**

Documente la técnica de modelado real que está por ser usado.

Presunciones del modelado

Muchas técnicas de modelado hacen presunciones específicas sobre los datos -por ejemplo, que todos los atributos tengan distribuciones uniformes, no encontrar valores no permitidos, el atributo de clase debe ser simbólico, etc. Registrar cualquiera de tales presunciones hechas.

4.2. Generación de la prueba de diseño

Tarea **Generar la prueba de diseño**

Antes de que nosotros en realidad construyamos un modelo, tenemos que generar un procedimiento o el mecanismo para probar la calidad y validez del modelo. Por ejemplo, en tareas de minería de datos supervisados como la clasificación, esto es común usar tasas de errores como medida de calidad para modelos de minería de datos. Por lo tanto, típicamente separamos el conjunto de datos en una serie y en un conjunto de prueba, construimos el modelo sobre el conjunto de series, y estimamos su calidad sobre el conjunto de prueba separado.

Salida **Prueba de diseño**

Describir el plan intencionado para el entrenamiento, la prueba, y la evaluación de los modelos. Un componente primario del plan determina como dividir un conjunto de datos disponible en datos de entrenamiento, datos de prueba, y conjunto de datos de validación.

4.3. Construcción del modelo

Tarea **Construir el modelo**

Ejecutar la herramienta de modelado sobre el conjunto de datos preparados para crear uno o más modelos.

Salidas **Parámetro de ajustes**

Con cualquier herramienta de modelado, hay a menudo un gran número de parámetros que pueden ser ajustados. Listar los parámetros y sus valores escogidos, también con el razonamiento para elegir los parámetros de ajustes.

Modelos

Estos son los modelos reales producidos por la herramienta de modelado, no un informe.

Descripciones del modelo

Describir los modelos obtenidos. Informar sobre la interpretación de los modelos y documentar cualquier dificultad encontrada con sus significados.

4.4. Evaluación del modelo

Tarea **Evaluar el modelo**

El ingeniero de minería de datos interpreta los modelos según su conocimiento de dominio, los criterios de éxitos de minería de datos, y el diseño de prueba deseado. El ingeniero de minería de datos juzga el éxito de la aplicación del modelado y descubre técnicas mas técnicamente; él se pone en contacto con analistas de negocio y expertos en el dominio luego para hablar de los resultados de la minería de datos en el contexto de negocio. Por favor note que esta tarea sólo se considera modelos, mientras que la fase de evaluación también toma en cuenta todos los otros resultados que fueron producidos en el curso del proyecto.

El ingeniero de minería de datos intenta clasificar los modelos. Él evalúa los modelos según los criterios de evaluación. Tanto como es posible, él también tiene en cuenta objetivos del negocio y criterios de éxito de negocio. En los grandes proyectos de minería de datos, el ingeniero de minería de datos aplica una sola técnica más de una vez, o genera resultados de minería de datos con varias técnicas diferentes. En esta tarea, él también compara todos los resultados según los criterios de evaluación.

Salida **Evaluación de modelos**

Resumir los resultados de esta tarea, listar las calidades de los modelos generados (por ejemplo, en términos de exactitud), y clasificar su calidad en relación con cada otro.

Parámetros de ajustes revisados

Según la evaluación del modelo, revise los parámetros de ajuste y témplelos para la siguiente corrida en la tarea de Construcción del Modelo. Repetir la construcción y evaluación del modelo hasta que crea que usted ha encontrado *el/los mejor/es m*odelo/s. Documentar todo como las revisiones y las evaluaciones.

5. Evaluación

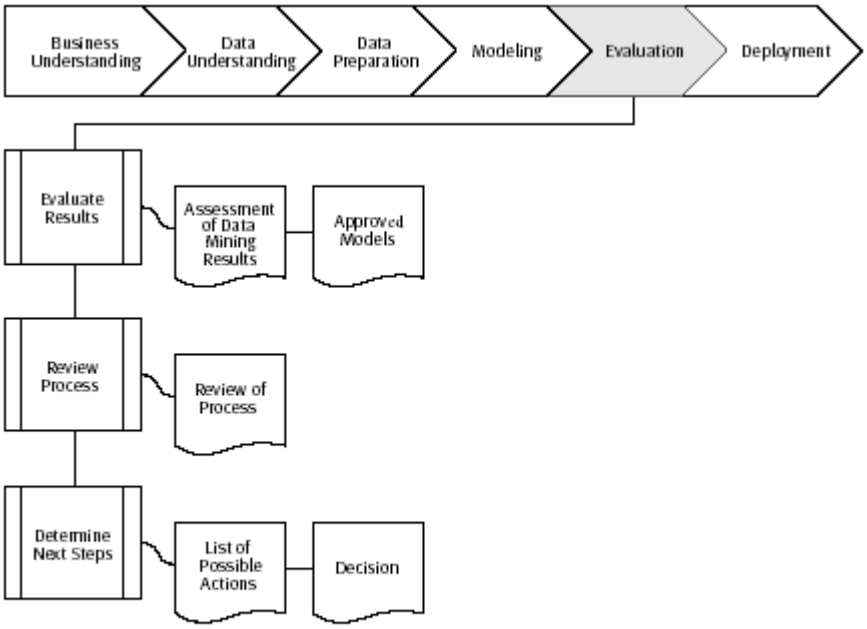

Figure 8: Evaluation

Figura 8: Evaluación

5.1. Evaluación de los resultados

Tarea **Evaluar los resultados**

Los pasos de la evaluación anterior trata con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.

Además, la evaluación también verifica otros resultados generados por la minería de datos. Los resultados de la minería de datos implican modelos que necesariamente son relacionados con los objetivos originales de negocio y todas los otros descubrimientos que no son relacionados necesariamente con los objetivos originales de negocio, pero también podría revelar desafíos adicionales, información, o insinuaciones para futuras direcciones.

Salida **Evaluación de los resultados de la minería de datos en lo que concierne a criterios de éxito de negocio**

Resumir los resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final en cuanto si el proyecto ya encuentra los objetivos iniciales de negocio.

Modelos aprobados

Después de la evaluación de modelos en lo que concierne a criterios de éxito de negocio, los modelos generados que encuentran los criterios seleccionados son los modelos aprobados.

5.2. Proceso de revisión

Tarea **Revisar el proceso**

En este punto, los modelos resultantes pasan a ser satisfactorios y a satisfacer las necesidades de negocio. Ahora es apropiado hacer una revisión más cuidadosa de los compromisos de la minería de datos para determinar si hay cualquier factor importante o tarea que de algún modo ha sido pasada por alto. Esta revisión también cubre cuestiones de calidad -por ejemplo: ¿Construimos correctamente el modelo? ¿Usamos sólo los atributos que nos permitieron usar y que están disponibles para análisis futuros?

Salida **Revisión de proceso**

Resumir la revisión de proceso y destacar las actividades que han sido omitidas y/o aquellas que deberían ser repetidas.

5.3. Determinación de los próximos pasos

Tarea **Determinar los próximos pasos**

Según los resultados de la evaluación y la revisión de proceso, el equipo de proyecto decide como proceder. El equipo decide si hay que terminar este proyecto y tomar medidas sobre el desarrollo si es apropiado, tanto iniciar más iteraciones, o comenzar nuevos proyectos de minería de datos. Esta tarea incluye los análisis de recursos restantes y del presupuesto, que puede influir en las decisiones.

Salida **Lista de posibles acciones**

Listar las acciones futuras potenciales, con los motivos a favor y en contra de cada opción.

Decisión

Describir la decisión en cuanto a como proceder, junto con el razonamiento.

6. Desarrollo

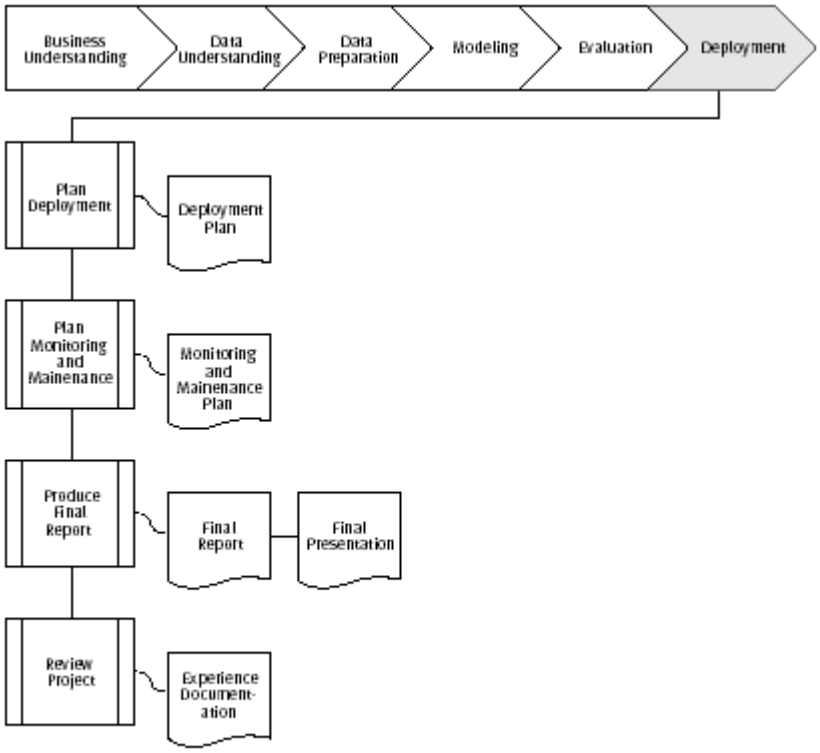

Figure 9: Deployment

Figura 9: Desarrollo

6.1. Desarrollo del plan

Tarea **Desarrollar el plan**

De acuerdo al desarrollo de los resultados de minería de datos en el negocio, esta tarea toma los resultados de la evaluación y determina una estrategia para el desarrollo. Si un procedimiento general ha sido identificado para crear el/los modelo/s relevante/s, este procedimiento es documentado aquí para el desarrollo posterior.

Salida **Desarrollo del plan**

Resumir la estrategia de desarrollo, incluyendo los pasos necesarios y como realizarlos.

6.2. Plan de supervisión y mantenimiento

Tarea **Planear la supervisión y el mantenimiento**

La supervisión y el mantenimiento son cuestiones importantes si los resultados de minería de datos son parte del negocio cotidiano y de su ambiente. La preparación cuidadosa de una estrategia de mantenimiento ayuda evitar largos periodos innecesarios de uso incorrecto de resultados de minería de datos. Para supervisar el desarrollo de los resultados de la minería de datos, el proyecto necesita un plan detallado de proceso de supervisión. Este plan tiene en cuenta el tipo específico de desarrollo.

Salida **Supervisión y plan de mantenimiento**

Resumir la estrategia de supervisión y mantenimiento incluyendo los pasos necesarios y como realizarlos.

6.3. Informe definitivo de producto

Tarea **Producir el informe final**

En el final del proyecto, el líder del proyecto y su equipo sobrescribe un informe final. Según el plan de desarrollo, este informe puede ser sólo un resumen del proyecto y sus experiencias (si estas aún no han sido documentadas como una actividad en curso) o esto puede ser una presentación final y comprensiva de los resultados de minería de datos.

Salidas **Informe definitivo**

Esto es el informe escrito final del compromiso de la minería de datos. Esto incluye todo el desarrollo anterior, el resumen y la organización de los resultados.

Presentación final

También a menudo habrá una reunión en la conclusión del proyecto en el que los resultados son presentados verbalmente al cliente.

6.4. Revisión del proyecto

Tarea **Revisar el proyecto**

Evaluar lo que fue correcto y lo que se equivocó, lo que fue bien hecho y lo que necesita para ser mejorado.

Salida **Documentación de la experiencia**

Resumir las experiencias importantes ganadas durante el proyecto. Por ejemplo, trampas, accesos engañosos, o las insinuaciones para seleccionar las mejores técnicas de minería de datos en situaciones similares podrían ser la parte de esta documentación. En proyectos ideales, la documentación de la experiencia también cubre cualquier informe que ha sido escrito por miembros individuales del proyecto durante las fases del proyecto y sus tareas.

III-La guía de usuario de CRISP-DM – AQUÍ ME QUEDE

1. Comprendiendo el negocio

1.1. Determinación de objetivos de negocio

Tarea **Determinar objetivos de negocio**

El primer objetivo del analista es comprender a fondo, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo el cliente tiene muchos objetivos y restricciones que compiten que deben ser correctamente equilibrados. El objetivo del analista debe destapar factores importantes en el principio del proyecto esto puede influir en el resultado final. Una consecuencia probable de descuidar este paso debe ser a expensas de un dar un gran esfuerzo de producir las respuestas correctas a las preguntas incorrectas.

Salida **Contexto**

Coteje la información que conoció sobre la situación de negocio de la organización al principio del proyecto. Estos detalles no sólo sirven para identificar mas estrechamente los objetivos de negocio a ser alcanzados, pero también sirven para identificar los recursos, tanto humano como material, que puede ser usado o sea necesario durante el curso del proyecto.

Actividades Organizar

- Desarrollar organigramas que identifiquen divisiones, departamentos, y grupos de proyectos. El organigrama debería también identificar los nombres de los gerentes y sus responsabilidades
- Identificar a personas claves en el negocio y sus roles
- Identificar a un patrocinador interno (el patrocinador financiero y el experto primario del dominio de usuario)
- Indicar si hay un comité de dirección y lista de miembros
- Identificar las unidades de negocio que son afectadas por el proyecto de minería de datos (por ejemplo, el Control de comercialización, Ventas, Finanzas)

Área del problema

- Identificar el área del problema (por ejemplo, el control de comercialización, el cuidado de cliente, el desarrollo comercial, etc.)
- Describir el problema en términos generales
- Comprobar el estado actual del proyecto (por ejemplo, Comprobar si ya esta claro que dentro de la unidad de negocio un proyecto de minería de datos debe ser realizado, o si la minería de datos necesita ser promovida como una tecnología clave en el negocio)
- Clarificar los requisitos previos del proyecto (por ejemplo, ¿Cuál es la motivación del proyecto? ¿La minería de datos ya está siendo usada en el negocio?)
- Si es necesario, preparar presentaciones y demostraciones de minería de datos para el negocio
- Identificar grupos de objetivos para el resultado de proyecto (por ejemplo, ¿Esperamos entregar un informe para la dirección superior o un sistema operacional para ser usado por usuarios finales inexpertos?)
- Identificar las necesidades de los usuarios y sus expectativas

Solución actual

- Describir cualquier solución usada actualmente para dirigir el problema
- Describen las ventajas y las desventajas de la solución corriente y el nivel al que esto es aceptado por los usuarios

Salida **Objetivos de negocio**

Describir el objetivo primario del cliente, desde una perspectiva de negocio. Además del objetivo de negocio primario, hay típicamente un gran número de preguntas relacionadas al negocio a las que al cliente le gustaría dirigir. Por ejemplo, el objetivo primario de negocio podría ser mantener a clientes actuales por predicción cuando ellos son propensos a moverse a un competidor, mientras un objetivo secundario de negocio podría ser de determinar si precios (comisiones) inferiores afectan sólo un segmento particular de clientes.

Actividades

- De manera informal describir el problema a ser solucionado
- Especificar todas las preguntas de negocio tan precisas como sea posible
- Especificar cualquier otras exigencias de negocio (por ejemplo, el negocio no quiere perder a ningún cliente)
- Especificar las ventajas esperadas en términos de negocio

¡Cuidado!

- Tener cuidado de establecer objetivos inalcanzables hechos por ellos tan realistas como posible.

Salida **Criterios de éxito de negocio**

Describir los criterios para un resultado exitoso o útil al proyecto desde el punto de vista del negocio. Esto podría ser bastante específico y fácilmente medible, como una reducción de cliente a un cierto grado, o general y subjetivo, como "dar ideas útiles en las relaciones". En el caso último, esté seguro de indicar quien haría el juicio subjetivo.

Actividades

- Especificar criterios de éxito de negocio (por ejemplo, Mejorar la tasa de respuesta en una campaña de correo en el 10 por ciento y marcar la tasa en el 20 por ciento)
- Identifican quien evalúa los criterios de éxito

¡Recuerde!

Cada uno de los criterios de éxito debería relacionarse con al menos uno de los objetivos especificados de negocio.

¡Buena Idea!

Antes del comienzo de la evaluación de situación, usted podría analizar las experiencias anteriores de este problema-

Internamente, usando CRISP-DM, o externamente, usando soluciones pre-empaquetadas.

1.2. Evaluación de la situación

Tarea **Evaluar la situación**

Esta tarea implica una investigación más detallada sobre todos los recursos, restricciones, presunciones, y otros factores que deberían ser considerados en la determinación del objetivo de análisis de datos y en el desarrollo del plan de proyecto.

Salida **Inventario de recursos**

Listar los recursos disponibles para el proyecto, incluyendo el personal (expertos de datos y de negocios, soportes técnicos, expertos en minería de datos), datos (extracciones fijas, acceso a datos existentes en almacenes de datos u operacionales), recursos computacionales (plataformas de hardware), y software (instrumentos de minería de datos, otros software relevantes).

Actividades Recursos de Hardware

- Identificar el hardware básico
- Establecer la disponibilidad del hardware básico para el proyecto de minería de datos
- Comprobar si la planificación del mantenimiento de hardware se opone a la disponibilidad del hardware para el proyecto de minería de datos.
- Identificar el hardware disponible para ser usado por la herramienta de minería de datos (si el instrumento es conocido en esta etapa)

Fuentes de datos y conocimientos

- Identificar las fuentes de datos
- Identificar el tipo de fuentes de datos (fuentes en línea, expertos, documentación escrita, etc.)
- Identificar fuentes de conocimiento
- Identificar el tipo de fuentes de conocimientos (fuentes en línea, expertos, documentación escrita, etc.)

- Comprobar herramientas disponibles y técnicas
- Describir el conocimiento de generalidades relevante (de manera informal o formalmente)

Fuentes de personal

- Identificar al patrocinador de proyecto (si difiere del patrocinador interno como en la Sección 1.1.1)
- Identificar al administrador de sistema, el administrador de base de datos, y el personal de soporte técnico para futuras preguntas
- Identificar a analistas de mercado, los expertos en minería de datos, y estadísticos, y comprobar su disponibilidad
- Comprobar la disponibilidad de expertos de dominio para fases posteriores

¡Recuerde!

Recuerde que el proyecto puede necesitar personal técnico en cualquier momento en todas partes del proyecto, por ejemplo durante la transformación de datos.

Salidas **Requerimientos, presunciones, y restricciones**

Listar todos los requerimientos del proyecto, incluyendo la planeación de la terminación, la comprensibilidad, y la calidad y seguridad de los resultados, así como cuestiones legales. Como la parte de esta salida, asegúrese de que le permiten usar los datos.

Listar las presunciones hechas por el proyecto. Estos pueden ser presunciones sobre los datos, que pueden ser verificados durante la minería de datos, pero también puede incluir presunciones no-comprobables relacionadas con el proyecto. Esto es en particular Importante de ponerlos en una lista si ellos afectarán la validez de los resultados.

Listar las restricciones hechas en el proyecto. Estas restricciones podrían implicar la carencia de recursos para terminar algunas tareas en el proyecto en el tiempo requerido, o allí pueden ser restricciones legales o éticas sobre el uso de los datos o la solución necesita terminar la tarea de minería de datos.

Actividades Requerimientos

- Especificar el perfil del grupo objetivo
- Capturar todas los requerimientos en la planificación
- Capturar los requerimientos de comprensibilidad, exactitud, desarrollar habilidades, mantenimiento, y repetibillidad del proyecto de minería de datos y los modelos resultantes.
- Capturar los requerimientos de seguridad, restricciones legales, de privacidad, información, y planificación de proyecto

Presunciones

- Aclarar todas las presunciones (incluyendo las implícitas) y las hechas por ellos explícitamente (por ejemplo, dirigir las cuestiones de negocio, a un número mínimo de clientes con la edad por encima de 50 es necesaria)
- Listar las presunciones sobre calidad de datos (por ejemplo, exactitud, disponibilidad)
- Listar las presunciones sobre factores externos (por ejemplo, cuestiones económicas, productos competitivos, avances técnicos)
- Aclarar presunciones que conducen a cualquiera de las estimaciones (por ejemplo, el precio de un instrumento específico es asumido para ser menor que 1,000 $)
- Listar todas las presunciones en cuanto a si es necesario entender y describir o explicar el modelo (Por ejemplo, como el modelo y los resultados son presentados a la dirección / patrocinador)

Restricciones

- Comprobar restricciones generales (por ejemplo, cuestiones legales, presupuesto, escalas de tiempo, y recursos)
- Comprobar el correcto acceso a fuentes de datos (por ejemplo, restricciones de acceso, la contraseña requerida)

- Comprobar la accesibilidad técnica de datos (los sistemas de operaciones, el sistema de administración de datos, el formato de archivo y de base de datos)
- Comprobar si el conocimiento relevante es accesible
- Comprobar restricciones de presupuesto (gastos fijos, gastos de implementación, etc.)

¡Recuerde!

La lista de presunciones también incluye presunciones al principio del proyecto, esto es, lo que el punto de inicio del proyecto ha sido.

Salidas **Riesgos y contingencias**

Listar los riesgos, es decir los acontecimientos que podrían ocurrir, impactando en la planificación, el costo, o el resultado. Listar los planes de contingencias respectivos: que acción será tomada para evitar o reducir al mínimo el impacto o recuperar de la ocurrencia de los riesgos previstos.

Actividades Identificar riesgos

- Identificar riesgos de negocio (por ejemplo, el competidor aparece primero con mejores resultados)
- Identificar riesgos de organización (por ejemplo, el departamento que solicita el proyecto no tiene financiación para el proyecto)
- Identificar riesgos financieros (por ejemplo, aumentar la financiación depende de los resultados iniciales de minería de datos)
- Identificar riesgos técnicos
- Identificar los riesgos que dependen de datos y de las fuentes de datos (por ejemplo, la mala calidad y cobertura)

Desarrollo de planes de contingencia

- Determinar condiciones en las que cada riesgo puede ocurrir
- Desarrollar planes de contingencia

Salida **Terminología**

Compilar un glosario de terminología relevante al proyecto. Esto debería incluir al menos dos componentes:

(1) Un glosario de terminología relevante de negocio, que forma parte de la comprensión de negocio disponible al proyecto

(2) Un glosario de terminología de minería de datos, ilustrada con ejemplos relevantes al problema de negocio en cuestión.

Actividades

- Comprobar la disponibilidad previa de glosarios; si no comience a bosquejar glosarios
- Hablar a expertos de dominio para entender su terminología
- Familiarizarse con la terminología de negocio

Salida **Costos y beneficios**

Preparar un análisis de costo-beneficio para el proyecto, comparando los gastos del proyecto con el beneficio potencial para el negocio si esto es exitoso

Actividades

- Estimar el costo para la colección de datos
- Estimar el costo de desarrollo y realización de una solución
- Identificar beneficios (por ejemplo, mejorar la satisfacción del cliente, ROI, y el aumento de las ganancias)
- Estimar gastos de operación

¡Buena Idea!

La comparación debería ser tan específica como sea posible, como esto permite un mejor caso de negocio para ser realizado.

¡Cuidado!

Acuérdese de identificar costos ocultos, como la extracción y preparación repetida de datos, cambios en los procesos laborales, y tiempo requerido para el entrenamiento.

1.3. Determinar objetivos de minería de datos

Tarea **Determinar objetivos de minería de datos**

Un objetivo de negocio declara objetivos en la terminología de negocio; un objetivo de minería de datos declara objetivos de proyecto en términos técnicos. Por ejemplo, el objetivo de negocio podría ser, “Aumentar la venta por catalogo a clientes existentes”, mientras un objetivo de minería de datos podría ser, “Predecir cuantas baratijas comprará un cliente, considerando sus compras durante los tres años pasados, información demográfica relevante, y el precio del artículo.”

Salidas **Objetivos de minería de datos**

Describir las salidas planeadas del proyecto que permiten el logro de los objetivos de negocio.

Note que estos son salidas normalmente técnicas.

Actividades

- Traducir las preguntas de negocio a objetivos de minería de datos (por ejemplo, una campaña de control de comercialización requiere la segmentación de clientes para decidir a quien acercarse en esta campaña; el nivel/tamaño de los segmentos debería ser especificado).
- Especificar datos tipo de problema de minería de datos (por ejemplo, la clasificación, la descripción, la predicción, y clustering). Para más detalles sobre tipos de problema de minería de datos, vea el Apéndice 2.

¡Buena idea!

Puede ser sabio redefinir el problema. Por ejemplo, modelar la retención de producto más que la retención del cliente cuando la retención del cliente entrega resultados muy tarde para afectar la salida.

Salida **Criterios de éxitos de minería de datos**

Definir los criterios para un resultado acertado para el proyecto en términos técnicos, por ejemplo un cierto grado de exactitud predictiva o un perfil de propensión-a-comprar con un nivel dado "elevación". Como con los criterios de éxitos del negocio, puede ser necesario describir estos en términos subjetivos, en el caso de que la persona o las personas que hacen el juicio subjetivo deberían ser identificadas.

Actividades

- Especificar los criterios para evaluar el modelo (por ejemplo, la exactitud del modelo, el funcionamiento y la complejidad)
- Definir el patrón de pruebas para los criterios de evaluación
- Especificar las reglas que dirigen criterios de evaluación subjetivos (por ejemplo, el habilidad de explicar del modelo y de los datos y la comprensión de mercadeo proporcionada por el modelo)

¡Tenga cuidado!

Recuerde que los datos que extraen criterios de éxito son diferentes a los criterios de éxito de negocio definidos antes.

Recuerde es sabio planear para el desarrollo desde el principio del proyecto.

1.4. Producción del plan del proyecto

Tarea **Producir el plan del proyecto**

Describir el plan propuesto para alcanzar los objetivos de minería de datos y así alcanzar de los objetivos de negocio.

Salida **Plan del Proyecto**

Listar las etapas para ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias. En cualquier parte donde posible, haga explícito las iteraciones en gran escala en el proceso de minería de datos- Por ejemplo, las repeticiones del modelado y fases de evaluación. Como parte del plan de proyecto, esto es también importante analizar dependencias entre el planeamiento de los tiempos y los riesgos. Marcar los resultados de estos análisis explícitamente

en el plan de proyecto, idealmente con acciones y recomendaciones para actuar si los riesgos son manifestados.

Aunque esto sea la única tarea en la que el plan de proyecto directamente es llamado, sin embargo debería ser consultado continuamente y repasado en todas partes del proyecto. Deberían consultar el plan de proyecto como mínimo siempre que una tarea nueva sea comenzada o una iteración futura de una tarea o una actividad esta comenzando.

Actividades

- Definir el plan de proceso inicial y hablar de la viabilidad con todo el personal incluido
- Combinar todos los objetivos identificados y técnicas seleccionadas en un procedimiento coherente que solucione las cuestiones del negocio y encuentre los criterios de éxito de negocio
- Estimar el esfuerzo y los recursos necesarios para alcanzar y desarrollar la solución. (Es útil considerar la experiencia de otras personas estimando escalas de tiempo para proyectos de minería de datos. Por ejemplo, es a menudo presumido que el 50-70 por ciento del tiempo y el esfuerzo en un proyecto de minería de es usado en la Fase de Preparación de Datos, mientras que solo un 20-30 por ciento es usado en la Fase de Comprensión de Datos, mientras que solo un 10-20 por ciento es gastado en cada uno de las Fase de Modelado, Evaluación, y Comprensión del Negocio Entendiendo y el 5-10 por ciento en la Fase de Desarrollo.)
- Identificar pasos críticos
- Marcar los puntos de decisión
- Marcar los puntos de revisión
- Identificar las principales iteraciones

Salida **Evaluación de Inicial de herramientas y técnicas**

Al final de la primera fase, el equipo de proyecto realiza una evaluación inicial de herramientas y técnicas. Aquí, es importante seleccionar una herramienta de minería de datos que soporte varios métodos para las diferentes etapas del proceso, ya que la selección de herramientas y técnicas puede influir en el proyecto entero.

Actividades

- Crear una lista de criterios de selección para herramientas y técnicas (o usar uno existente si está disponible)
- Escoger herramientas y técnicas posibles
- Evaluar la adecuación de técnicas
- Revisar y priorizar técnicas aplicables según la evaluación de soluciones alternativas

2. Comprensión de Datos

2.1. Recolección de datos iniciales

Tarea **Recoger datos iniciales**

Obtener los datos (o el acceso a los datos) listados en los recursos de proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de datos. Por ejemplo, si usted tiene la intención de usar una herramienta específica para comprender los datos, es lógico cargar sus datos en esta herramienta.

Salida **Informe de la recolección de datos inicial**

Describir toda la variedad de datos usados para el proyecto, e incluya cualquier requerimiento de selección para datos más detallados. El informe de colección de datos también debería definir si algunos atributos son relativamente más importantes que otros.

Recuerde que cualquier evaluación de calidad de datos debería ser hecha no solamente de las fuentes de datos individuales, pero también de algunos datos que son resultado de fuentes de datos que se combinan. Por inconsistencias entre las fuentes, los datos combinados pueden presentar los problemas que no existen en las fuentes de datos individuales.

Actividades **Planificación de requerimientos de datos**

Planee que información es necesaria (por ejemplo, sólo para atributos determinados, o la información adicional específica)

Comprobar si toda la información necesaria (para resolver los objetivos de la minería de datos) esta en realidad disponible

Criterios de selección

- Especificar los criterios de selección (por ejemplo, ¿Qué atributos son necesarios para los objetivos específicos de minería de datos? ¿Que atributos han sido identificados como no pertinentes? ¿Cuantos atributos podemos manejar con las técnicas escogidas?)
- Elegir tablas/archivos de interés
- Elegir datos dentro de una tabla/archivo
- Pensar cuanto tiempo de una historial habría que usar (por ejemplo, si 18 meses de datos están disponibles, sólo 12 meses pueden ser necesarios para el ejercicio)

¡Tenga cuidado!

Estar consciente de que los datos recolectados de diferentes fuentes pueden dar lugar a problemas de calidad cuando sean combinados (Por ejemplo, los archivos de dirección combinados con una base de datos de cliente pueden mostrar inconsistencias de formato, invalidez de datos, etc.).

Inserción de datos

- Si los datos contienen libre entradas de texto, ¿tenemos que codificarlos para modelar o necesitamos agruparlos en entradas específicas?
- ¿Cómo podemos encontrar atributos omitidos?
- ¿Cómo podemos mejorar la extracción los datos?

¡Buena Idea!

Recordar que algún conocimiento sobre los datos puede estar disponible de fuentes no-electrónicas (Por ejemplo, de gente, de texto impreso, etc.).

Recordar que puede ser necesario a preproceso de los datos (datos de serie tiempo, promedios ponderados, etc.).

2.2. Descripción de datos

Tarea **Describir datos**

Examine las propiedades "gruesas" de los datos obtenidos y el informe sobre los resultados.

Salida **Informe de descripción de datos**

Descripción de los datos que han sido obtenidos, incluyendo el formato de los datos, la cantidad de los datos

(Por ejemplo, el número de registros y campos internos de cada tabla), las identidades de los campos, y cualquier otro rasgo superficial que haya sido descubierto.

Actividades Análisis Volumétrico de datos

- Identificar datos y métodos de captura
- Acceder a las fuentes de datos
- Usar análisis estadísticos si es apropiado
- Reportar las tablas y sus relaciones
- Compruebe el volumen de datos, el número de múltiplos, la complejidad
- Notar si los datos contienen entradas de texto libres

Atributo tipos y valores

- Comprobar la accesibilidad y disponibilidad de atributos
- Comprobar los tipos de atributos (numérico, simbólico, la taxonomía, etc.)
- Comprobar el rango de valores de los atributos
- Analizar los atributos correlativos (correlaciones de atributo)
- Comprender el significado de cada atributo y clasificar (describir) el valor en términos de negocio

- Para cada atributo, calcular la estadística básica (por ejemplo, calcule la distribución, el promedio, el máximo, el mínimo, la desviación estándar, la varianza, la moda, la inclinación, etc.)
- Analizar la estadística básica y relacionan los resultados con su significado en términos de negocio
- Decidir si el atributo es relevante para los objetivos específicos de la minería de datos
- Determinar si el significado del atributo es usado coherentemente (conscientemente)
- Entrevistar a expertos de dominio para obtener su opinión sobre la importancia de los atributos
- Decidir si es necesario equilibrar los datos (basado en las técnicas que modelan a ser usado)

Claves

- Analizar relaciones claves
- Comprobar la cantidad de coincidencias entre valores de atributos claves a través de tablas

Revisión de Objetivos/Presunciones

- Actualizar la lista de presunciones, si es necesario

2.3. Exploración de datos

Tarea **Explorar datos**

Esta tarea aborda las preguntas de minería de datos que pueden ser dirigidas usando la interrogación, la visualización, y técnicas de informe. Estos análisis pueden directamente dirigir los objetivos de minería de datos. Sin embargo, ellos pueden también contribuir a refinar la descripción de datos e informes de calidad, y alimentar internamente la transformación y otros pasos de preparación de datos necesario antes de que pueda ocurrir un futuro análisis.

Salida **Informe de exploración de datos**

Describir los resultados de esta tarea, incluyendo las primeras conclusiones o las hipótesis iniciales y su impacto sobre el resto del proyecto. El informe también puede incluir gráficos y diseños (plots) que indican las características de los datos o los puntos de interés de subconjuntos de datos dignos de una futura investigación.

Actividades Exploración de Datos

- Analizar en detalles las propiedades de atributos interesantes (por ejemplo, la estadística básica, las sub-poblaciones interesantes)
- Identificar las características de las sub-poblaciones

Formar suposiciones para análisis futuro

- Considerar y evalúan la información y conclusiones en el informe de descripciones de datos
- Formar una hipótesis e identifican acciones
- Transforman la hipótesis en un objetivo de minería de datos, si es posible
- Aclarar objetivos de minería de datos o hacerlos más exactos. Una búsqueda "ciega" no es necesariamente inútil, pero una búsqueda más dirigida hacia objetivos de negocio es preferible.
- Realizar un análisis básico para verificar la hipótesis

2.4. Verificación de la calidad de datos

Tarea **Verificar la calidad de datos**

Examine la calidad de los datos, dirigiendo preguntas como: Es los datos completos (¿esto cubre todos los casos requeridos?) ¿Hay en ellos errores o ellos contienen errores? ¿Si hay errores, como son ellos? ¿Hay valores omitidos en los datos? Si es así, ¿cómo son representados, donde ocurren, y como son ellos?

Salida **Informe de calidad de datos**

Listar los resultados de la verificación de calidad de datos; si hay problemas de calidad, Listar las posibles soluciones.

Actividades

Identificar valores especiales y catalogar su significado

Revisión de atributos claves

- Comprobar la cobertura (por ejemplo, si todos los valores posibles son representados)
- Comprobar las claves
- Verificar que los significados de los atributos y valores contenidos se satisfacen simultáneamente
- Identificar atributos omitidos y campos en blanco
- Establecer el significado de datos que faltan o fallan
- Comprobar los atriibutos con los valores diferentes que tienen significados similares (por ejemplo, la grasa baja, la dieta)
- Comprobar la ortografía y el formato de valores (por ejemplo, mismo valor pero a veces comienza con una letra minúscula, a veces con una letra mayúscula)
- Comprobar las desviaciones, y deciden si una desviación es "ruido" o puede indicar un fenómeno interesante
- Comprobar la plausibilidad de valores, (por ejemplo, todos los campos que tienen el mismo o casi los mismos valores)

¡Buena idea!

Repasar cualquiera de los atributos que dan respuestas que están en desacuerdo con el sentido común (por ejemplo, adolescentes con altos niveles de ingreso).

Use plots de visualización, histogramas, etc. para revelar inconsistencias en los datos.

Calidad de datos en archivos planos

- Si los datos son almacenados en archivos planos, comprobar que delimitador es usado y si esto es usado coherentemente en todos los atributos
- Si los datos son almacenados en archivos planos, comprobar el número de campos en cada registro para ver si ellos coinciden

Ruido e inconsistencias entre fuentes

- Comprobar consistencia y superabundancia entre fuentes diferentes
- Planear para tratar el ruido
- Descubrir el tipo de ruido y que atributos son afectados

¡Buena idea!

Recuerde que puede ser necesario excluir algunos datos ya que ellos no exponen comportamiento positivo o negativo (por ejemplo, al comprobar en el comportamiento del préstamo de clientes, excluye a todo los que nunca han tomado prestado, aquellos que no financian una hipoteca de casa, aquellos cuya hipoteca se acerca a la madurez, etc.).

Revisar si las presunciones son válidas o no, considerando la información real o actual en los datos y el conocimiento de negocio.

3. Preparación de los datos

Salida **Conjunto de datos**

Estos son los conjuntos de dato(s) producidos por la fase de preparación de datos, usada para modelar o para el trabajo de análisis principal del proyecto.

Salida **Descripción del conjunto de datos**

Esto es la descripción del conjunto de datos(s) usado para el modelado o para el trabajo de análisis principal del proyecto.

3.1. Datos seleccionados

Tarea **Seleccionar datos**

Decidir los datos a ser usados para el análisis. Los criterios incluyen la importancia a los objetivos de minería de datos, la calidad, y las restricciones técnicas como los límites en el volumen de datos o en los tipos de datos.

Salida **Razonamiento para inclusión/exclusión**

Listar los datos a ser usados / excluidos y los motivos para estas decisiones.

Actividades

- Recogen datos adicionales apropiados (de diferentes fuentes - internos así como externos)
- Realizar las pruebas de importancia y correlación para decidir si los campos son incluidos
- Reconsideran Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de calidad de los datos y en la exploración de datos (esto es, puede desear incluir/excluir otros juegos de datos)
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de experiencia de modelado (esto es, la evaluación del modelo puede mostrar que otros conjuntos de datos son necesarios)
- Seleccionar diferentes subconjuntos de datos (por ejemplo, atributos diferentes, sólo los datos que encuentran ciertas condiciones)
- Considerar el uso de técnicas de muestreo (por ejemplo, una solución rápida puede implicar la prueba dura y el entrenamiento del conjunto de datos o la reducción del tamaño de la conjunto de datos de prueba, si la herramienta no puede manejar conjunto de datos llenos. Esto puede también ser útil para tener muestras ponderadas para dar la distinta importancia a atributos diferentes o valores diferentes del mismo atributo.)
- Documentar el razonamiento para la inclusión/exclusión
- Comprobar técnicas disponibles para el muestreo de datos

¡Buena idea!

Basado en Criterios de Selección de Datos, decidir si uno o más atributos son más importantes que otros el correspondiente peso de los atributos. Decidir, basado en el contexto (esto es, el uso, la herramienta, etc.), como manejarse con el peso.

3.2. Limpieza de datos

Tarea **Limpiar datos**

Elevar la calidad de datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de los datos, la inserción de faltas apropiadas, o técnicas más ambiciosas como la estimación de datos omitidos por modelado.

Salida **Informe de la limpieza de datos**

Describir las decisiones y las acciones que fueron tomados para dirigir los problemas de calidad de datos informados durante la Tarea de Verificación de Calidad de Datos. Si los datos están para ser usados en el ejercicio de minería de datos, el informe debería dirigir cuestiones de calidad de datos excepcionales y el efecto posible que esto podría tener sobre los resultados.

Actividades

- Reconsiderar como tratar con cualquier tipo de ruido observado
- Corregir, remover, o ignorar el ruido
- Decidir como tratar con valores especiales y su significado. El área de valores especiales puede dar lugar a muchos resultados extraños y con cuidado deberían ser examinados. Los ejemplos de valores especiales podrían surgir por los resultados tomados de una revisión donde algunas cuestiones no fueron preguntadas o no fueron contestadas. Esto podría terminar en un valor de 99 para datos desconocidos. Por ejemplo, 99 para estado civil o afiliación política. Los valores especiales también podría surgir cuando los datos son truncados por ejemplo., 00 para gente de 100 años o para todos los coches con 100,000 kilómetros en el odómetro.
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de los datos limpiados (esto es, usted puede desea incluir/excluir otros conjuntos de datos).

¡Buena idea!

Recuerde que algunos campos pueden ser irrelevantes a los objetivos de minería de datos y, por lo tanto, el ruido en aquellos campos no tiene ninguna importancia. Sin embargo, si el ruido es ignorado por estos motivos, esto debería ser totalmente documentado como circunstancias que pueden cambiarse más tarde.

3.3. Construcción de datos

Tarea **Construir datos**

Esta tarea incluye la construir de operaciones de preparación de datos tales como la producción de atributos derivados, completar registros nuevos, o transformar valores para atributos existentes.

Actividades

- Comprobar los mecanismos de construcción disponibles con la lista de herramientas sugeridas para el proyecto
- Decidir si esto es lo mejor para realizar la construcción dentro de la herramienta o fuera de ella (esto es, que es más eficiente, exacto, repetible)
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de construcción de datos (esto es, usted puede desear incluir/excluir otros conjuntos de datos)

Salida **Atributos derivados**

Los atributos derivados son los atributos nuevos que son construidos de uno o atributos más existentes en el mismo registro. Un ejemplo podría ser: área = longitud * anchura.

¿Por qué deberíamos tener que construir atributos derivados durante el curso de una investigación de minería de datos? No debería pensarse que sólo los datos de bases de datos u otras fuentes deberían ser usados en la construcción de un modelo. Los atributos derivados podrían ser construidos porque:

- El conocimiento del contexto nos convence que algún hecho es importante y debería ser representado aunque no tengamos ningún atributo actualmente para representarlo
- El algoritmo de modelado en uso maneja los sólo ciertos tipos de datos -por ejemplo estamos usando regresión lineal y sospechamos que hay ciertas no-linealidades que serán incluidos en el modelo
- El resultado de la fase de modelado sugiere que ciertos hechos no sean cubiertos

Actividades Derivar atributos

- Decidir si cualquier atributo puede ser normalizado (por ejemplo, usando un algoritmo de agrupamiento (clustering) con el periodo y el ingreso, en ciertas divisas, el ingreso se controlará)
- Considerar agregar nueva información sobre la importancia relevante de los atributos para agregar de nuevos atributos (Por ejemplo, atributos con peso, normalización ponderada)
- ¿Cómo se puede construir o imputar atributos faltantes? [Decidir el tipo de construcción (por ejemplo, la combinación, el promedio, la inducción).]
- Agregar atributos nuevos a los datos acceso de acceso

¡Buena idea!

Antes de agregar Atributos Derivados, intente determinar si y como ellos facilitan el proceso de modelado o facilitan el algoritmo de modelado. Quizás “el ingreso por persona” es un mejor/más fácil atributo para usar que “el ingreso por casa.” No saque atributos simplemente para reducir el número de atributos de entrada.

Otro tipo de atributo derivado es la transformación de un atributo individual, por lo general realizado para cubrir las necesidades de las herramientas de modelado.

Actividades Transformaciones de atributo individual

- Especificar los pasos de transformaciones necesarias en los términos de facilitar las transformación disponibles (por ejemplo, cambiar un binning de un atributo numérico)
- Realizar pasos de transformación

¡Buena idea!

Las transformaciones pueden ser necesarias para cambiar rangos a campos simbólicos (por ejemplo, años a rangos de edad) o campos simbólicos (“definitivamente sí”, “sí”, “no se sabe,” "no") a valores numéricos. Las herramientas de modelado o los algoritmos a menudo los requieren.

Salida **Registros generados**

Los registros generados son registros completamente nuevos, que agregan nuevo conocimiento o representan nuevos datos que de otro modo no son representado (por ejemplo, habiendo segmentado los datos, puede ser útil generar un registro para represente al miembro prototípico de cada segmento para un tratamiento futuro).

Actividades

Comprobar por técnicas disponibles si es necesario (por ejemplo, mecanismos para construir prototipos para cada segmento de datos segmentados).

3.4. Integración de datos

Tarea **Integrar datos**

Estos son métodos para combinar la información de múltiples tablas u otras fuentes de información para crear nuevos registros o valores.

Salida **Datos combinados**

La combinación de tablas se refiere a la unión de dos o más tablas que tienen diferente información sobre los mismos objetos. En esta etapa, también puede ser aconsejable generar registros nuevos. También puede ser recomendado para generar valores agregados.

La agregación se refiere a operaciones donde los nuevos valores son calculados por información resumida de múltiples registros y/o tablas.

Actividades

- Comprobar si las aplicaciones de integración son capaces de integrar las fuentes de entrada como se requiere
- Integrar fuentes y resultados almacenados
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de integración de datos (esto es, usted puede desear incluir/excluir otros conjuntos de datos)

¡Buena idea!

Recordar que algún conocimiento puede estar contenido en el formato no-electrónico.

3.5. Formateo de datos

Tarea **Formatear datos**

Transformar formateando se refiere principalmente a modificaciones sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.

Salida **Datos reformateados**

Algunas herramientas tienen requerimientos sobre la orden de los atributos, tal que el primer campo sea un único identificador para cada registro o el campo último ser el juego de resultados que el modelo debe predecir.

Actividades Atributos reorganizados

Algunas herramientas tienen requerimientos sobre la orden de los atributos, tal que el primer campo sea un único identificador para cada registro o el campo último ser el juego de resultados que el modelo debe predecir.

Reordenando registros

Podría ser importante cambiar el orden de los registros en el conjunto de datos. Quizás el instrumento de modelado requiere que los registros sean clasificados según el valor del atributo de resultado.

Reformateado valores internos

- Estos son cambios puramente sintácticos hechos para satisfacer las exigencias de la herramienta específica de modelado
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de limpieza de datos (esto es, usted puede desear incluir/excluir otros conjuntos de datos)

4. Modelado

4.1. Seleccionar técnicas de modelado

Tarea **Seleccionar técnicas de modelado**

Como el primero paso en modelado, seleccionar la técnica de modelado inicial actual. Si múltiples esta para ser aplicados, realizar separadamente esta tarea para cada técnica.

Recuerde que no todos los instrumentos y técnicas son aplicables a toda y cada tarea. Para ciertos problemas, sólo algunas técnicas son apropiadas (Vea el Apéndice 2, donde las técnicas asignan para ciertos tipos de problemas de minería de datos es hablada más detalladamente).

"Requerimientos políticos" y otras restricciones adicionales limitan las opciones disponibles para el ingeniero de minería de datos. Puede ser solo una herramienta o técnica están disponibles para solucionar el problema a mano - y que el instrumento no pueda ser absolutamente lo mejor, de un punto de vista técnico.

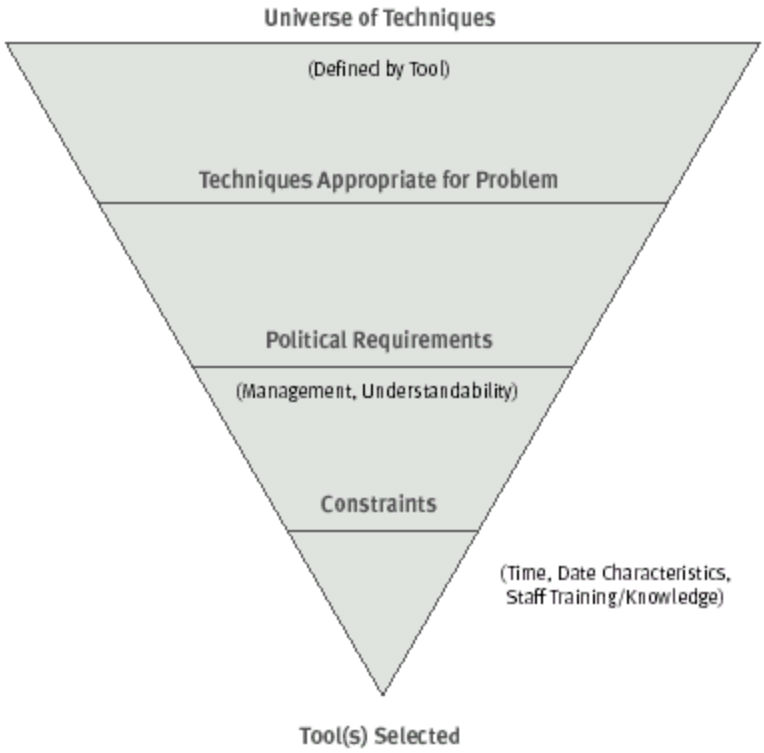

Figure 10:
Universe of Techniques

Figura 10: Universidad (Diversidad) de Técnicas

Salida **Técnicas de modelado**

Registrar las técnicas de modelado real que se usa.

Actividades

Decidir las técnicas apropiada para el ejercicio, teniendo en cuenta la herramienta seleccionada.

Salida **Presunciones de modelado**

Muchas técnicas de modelado realizan presunciones específicas sobre los datos.

Actividades

- Definir cualquier presunciones construidas hechas por la técnica sobre los datos (por ejemplo, la calidad, el formato, la distribución)
- Comparar estas presunciones con aquellas de el Informe de Descripción de Datos
- Asegurarse que estas presunciones sostienen y volver a la Fase de Preparación de Datos, si es necesario

4.2. Generar el diseño de prueba

Tarea **Generar el diseño de prueba**

Antes de construir un modelo, es necesario definir un procedimiento para probar la calidad del modelo y la validez. Por ejemplo, en tareas de minería de datos supervisadas como la clasificación, es común usar tasas de error como medidas de calidad para modelos de minería de datos. Por lo tanto, el diseño de prueba especifica que el conjunto de datos debería ser separado en el entrenamiento y en el conjunto de prueba. El modelo esta construido sobre el conjunto de entrenamiento y su calidad estimada sobre el conjunto de prueba.

Salida **Diseño de Prueba**

Describir el plan deliberado para el entrenamiento, las pruebas, y la evaluación de los modelos. Un componente primario del plan es para decidir como dividir el conjunto de datos disponible sobre datos que se entrenan, datos de prueba, y conjunto de pruebas de validación.

Actividades

- Comprobar que existe diseños de prueba separadamente para cada objetivo de minería de datos
- Decidir los pasos necesarios (el número de iteraciones, el número de desviaciones o curvas, etc.)
- Preparar los datos requeridos para la prueba

4.3. Construcción del modelo

Tarea **Construir el modelo**

Correr la herramienta de modelado sobre el conjunto de datos listo para crear uno o más modelos.

Salida **Parámetros de ajuste**

Con cualquier herramienta de modelado, hay a menudo un gran número de parámetros que pueden ser ajustados. Listar los parámetros y sus valores seleccionados, con la explicación (el razonamiento) para la elección.

Actividades

- Determinar los parámetros iniciales
- Documentar las razones para elegir aquellos valores

Salida **Modelos**

Controle la herramienta de modelado en el conjunto de datos listos para crear uno o más modelos.

Actividades

- Ejecutar la técnica seleccionada sobre el conjunto de datos de entrada para producir el modelo
- Post-procesar los resultados de minería de datos (por ejemplo, editar reglas, mostrar árboles)

Salida **Descripción del modelo**

Describir el resultado del modelado y evaluar su exactitud esperada, la robustez, y defectos posibles.

Informar sobre la interpretación de los modelos y encontrar cualquier de las dificultades.

Actividades

- Describir cualquier características del modelo actual que puede ser útil para el futuro
- Ajustar parámetro de entorno (de registro) usado para producir el modelo
- Dar una descripción detallada del modelo y cualquier rasgo especial
- Para modelos basados por regla, listar las reglas producidas, más cualquier evaluación de cada-regla o la exactitud y alcance total del modelo
- Para modelos no transparentes, listar cualquier información técnica sobre el modelo (como la topología de las redes neuronales) y cualquier descripción de comportamiento producido por el proceso de modelado (como la exactitud o la sensibilidad)
- Describir el comportamiento del modelo y la interpretación
- Expresar conclusiones respecto a los patrones en los datos (si hay alguno); a veces el modelo revela hechos importantes sobre los datos sin un proceso de evaluación separado (por ejemplo, que la salida o la conclusión son duplicadas en una de las entradas)

4.4. Evaluación del modelo

Tarea **Evaluar el modelo**

El modelo ahora debería ser evaluado para asegurar que se encontró criterios de éxito de la minería de datos y aprobar los criterios de prueba deseados. Esto es una evaluación puramente técnica basada en el resultado de las tareas modelado.

Salida **Evaluación del modelo**

Resumir los resultados de esta tarea, listar las calidades de los modelos generados (por ejemplo, en términos de exactitud), y el nivel de su calidad en relación a cada otro.

Actividades

- Evaluar los resultados en lo que concierne a criterios de evaluación
- Probar los resultados según una estrategia de prueba (por ejemplo: Corrida y Prueba, Validación cruzada, bootstrapping, etc.)
- Comparar los resultados de la evaluación y la interpretación
- Crear la clasificación de resultados en lo que concierne a criterios de éxito y evaluación
- Seleccionar los mejores modelos
- Interpretar los resultados en términos de negocio (tanto como sea posible en esta etapa)
- Conseguir comentarios de los modelos por expertos en datos o en el dominio
- Chequear la credibilidad del modelo
- Comprobar los efectos sobre los objetivos de minería de datos
- Comprobar los modelan contra una base de conocimiento determinada para ver si la información descubierta es nueva y útil
- Comprobar la fiabilidad de los resultados
- Analizar el potencial para el desarrollo de cada resultado
- Si hay una descripción verbal del modelo generado (por ejemplo, en forma de reglas), evaluar las reglas: ¿Ellos son lógicos, o ellos son factibles, hay demasiadas reglas o hay demasiado poco, ellos violan el sentido común?
- Evaluar resultados
- Conseguir ideas específicas de cada técnica de modelado y ciertos parámetros de ajustes que conduzcan a resultados buenos/malos

¡Buena idea!

"Tablas de Elevación" y "Tablas de Beneficio" pueden ser construidas para determinar los bien que el modelo predice.

Salida **Revisación de parámetros de ajuste**

Según la evaluación del modelo, revise parámetros de ajuste y témplelos para la siguiente corrida en la tarea de Construcción del Modelo. Itere (repita) la construcción del modelo y evalúe hasta que usted encuentre el mejor modelo.

Actividades

Ajustar parámetros para producir mejores modelos.

5. Evaluación

Los pasos de evaluación previa tratan con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado que el modelo encuentra los objetivos de negocio, y procura determinar si hay alguna razón de negocio por qué este modelo sea deficiente. Esto compara resultados con los criterios de evaluación definidos en el principio del proyecto.

Un modo bueno de definir las salidas totales de un proyecto de minería de datos es usar la ecuación:

RESULTADOS = MODELOS + CONCLUSIONES

En esta ecuación, definimos que la salida total del proyecto de minería de datos no es solamente los modelos (aunque ellos sean, desde luego, importantes) pero también las conclusiones, las que definimos como algo (aparte del modelo) que es importante en

- la búsqueda de los objetivos de negocio o importante para arribar a nuevas preguntas,
- las líneas de aproximación, o
- los efectos negativos (por ejemplo, los problemas de calidad de datos descubierto por el uso de la minería de datos).

Notar: Aunque el modelo esté directamente conectado a las preguntas de negocio, las conclusiones no necesariamente están relacionadas con cualquiera de las preguntas u objetivos, mientras ellos son importantes para el promotor del proyecto.

5.1. Evaluación de los resultados

Tarea **Evaluar los resultados**

Este paso evalúa el grado al que el modelo encuentra los objetivos de negocio, y procura determinar si hay alguna razón de negocio por el cual este modelo es deficiente. Otra opción es probar el (los) modelo(s) sobre la aplicación de prueba en el sistema verdadero, si permiten las restricciones de tiempo y de presupuesto.

Además, la evaluación también evalúa otros resultados generados por la minería de datos. Los resultados de minería de datos cubren los modelos que están relacionados con los objetivos originales de negocio y todas las demás conclusiones. Unos son relacionados con los objetivos de negocios originales mientras que otros podrían revelar desafíos adicionales, información, o ideas para futuras administraciones (direcciones).

Salida **Evaluación de los resultados de minería de datos en lo que respecta a criterios de éxito de negocio**

Resumir resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final relacionada a si el proyecto ya encuentra los objetivos iniciales de negocio.

Actividades

- Comprender los resultados de la minería de datos
- Interpretar los resultados en términos de la aplicación (del uso)
- Comprobar efectos sobre los objetivos de minería de datos
- Comprobar los resultados de minería de datos contra la base de un conocimiento determinado para ver si la información descubierta es nueva y útil
- Evaluar y estimar los resultados en lo que respecta a criterios de éxito de negocio (esto es, el proyecto ha alcanzado los Objetivos de Negocio originales)
- Comparar los resultados de la evaluación y la interpretación
- Clasificar los resultados en lo que respecta a criterios de éxito de negocio
- Comprobar el efecto de los resultados sobre el objetivo (fin) de la aplicación inicial
- Determinar si hay nuevos objetivos de negocio para ser dirigidos más tarde en el proyecto, o en nuevos proyectos
- Expresar recomendaciones para proyectos futuros de minería de datos

Salida **Modelos aprobados**

Después de evaluar los modelos con respecto a los criterios de éxito de negocio, seleccionar y aprobar los modelos generados que encontraron los criterios seleccionados.

5.2. Proceso de revisión

Tarea **Revisar el proceso**

En este punto, el modelo resultante parece ser satisfactorio y parece satisfacer necesidades de negocio. Es ahora apropiado hacer una revisión más cuidadosa de las promesas de minería de datos para determinar si hay algún factor importante o tarea que de algún modo ha sido pasada por alto. En esta etapa del ejercicio de minería de datos, el Proceso de Revisión toma la forma de una Revisión de Garantía de Calidad.

Salida **Revisión de procesos**

Resumir el proceso de revisión y poner en una lista las actividades que han sido omitidas y/o deberían ser repetidas.

Actividades

- Proporcionar una descripción del proceso de minería de datos usado
- Analizar el proceso de minería de datos. Para cada etapa del proceso pregunte:
    - ¿Esto fue necesario?
    - ¿Esto fue ejecutado óptimamente?
    - ¿En que modo podría ser mejorado?
- Identificar fracasos

- Identificar pasos desviados (de engaños)
- Identificar acciones alternativas posibles y/o caminos inesperados en el proceso
- Revisar resultados de minería de datos en lo que concierne a criterios de éxito de negocio

5.3. Determinación de los próximos pasos

Tarea **Determinar los próximos pasos**

Basado en los resultados de evaluación y la revisión de proceso, el equipo de proyecto decide como proceder.

Las decisiones a ser hechas incluyen si hay que terminar este proyecto y seguir adelante al desarrollo, para iniciar futuras Iteraciones, o establecer nuevos proyectos de minería de datos.

Salida **Lista de acciones posibles**

Lista acciones futuras posibles con los motivos para y contra de cada opción.

Actividades

- Analizar e potencial para el desarrollo de cada resultado
- Estimar el potencial para la mejora de proceso actual
- Comprobar los recursos restantes para determinar si ellos permiten iteraciones de proceso adicionales (o si recursos adicionales pueden estar siendo disponibles)
- Recomendar continuar con las alternativas
- Refinar el plan de proceso

Salida **Decisión**

Describir las decisiones hechas, con el razonamiento para ello.

Actividades

- Clasificar las acciones posibles
- Seleccionar una de las acciones posibles
- Documentar las razones para la elección

6. Desarrollo

6.1. Plan de desarrollo

Tarea **Desarrollo del Plan**

Esta tarea comienza con la evaluación de los resultados y concluye con una estrategia para el desarrollo de los resultados de la minería de datos en el negocio.

Salida **Plan de Desarrollo**

Resumir la estrategia de desarrollo, incluyendo los pasos necesarios y como realizarlos.

Actividades

- Resumir resultados desarrollados
- Construir y evaluar los planes alternativos para el desarrollo
- Decidir para cada resultado de conocimiento o información distinto
- Determinar como el conocimiento o la información serán propagados (generados) a los usuarios
- Decidir como será supervisado el uso del resultado y medido sus beneficios (donde sea aplicable)
- Decidir por cada resultado de modelo desarrollado o de software
- Establecer como el modelo o el resultado de software serán desplegados dentro de los sistemas de la organización
- Determinar como su empleo será supervisado y medido sus beneficios (donde sea aplicable)
- Identificar posibles problemas durante el desarrollo (peligros a ser evitados)

6.2. Supervisión y mantenimiento del plan

Tarea **Supervisar y mantener el plan**

La supervisión y el mantenimiento son cuestiones importantes si los resultados de la minería de datos se hacen parte del negocio cotidiano y de su ambiente. Una preparación cuidadosa de una estrategia de mantenimiento ayuda evitar innecesariamente largos períodos de uso incorrecto de los resultados de minería de datos. Para supervisar el desarrollo de los resultados de minería de datos, el proyecto necesita un plan detallado para supervisar y mantener. Este plan tiene en cuenta el tipo específico de desarrollo.

Salida **Plan de supervisión y mantenimiento**

Resumir la estrategia de supervisión y mantenimiento, la inclusión de pasos necesarios y como realizarlos.

Actividades

- Comprobar aspectos dinámicos (esto es, ¿qué cosas podrían cambiar en el entorno?)
- Decidir como será supervisada la precisión
- Determinar cuando el resultado de minería de datos o el modelo no deberían ser usados más. Identifique criterios (la validez, el límite de la exactitud, nuevos datos, cambios en el dominio de aplicación, etc.), y que debería pasar si el modelo o el resultado no pueden ser más usados. (Actualización del modelo, establecimiento de nuevos proyectos de minería de datos, etc.).
- ¿Cambiarán con el tiempo los objetivos de negocio del uso empleo del modelo? Documentar totalmente el problema inicial que el modelo intentaba solucionar.
- Desarrollar el plan de mantenimiento y la supervisión.

6.3. Producción de Informe definitivo

Tarea **Producir Informe definitivo**

En el final del proyecto, el equipo de proyecto sobrescribe un informe definitivo. Según el plan de desarrollo, este informe puede ser sólo un resumen del proyecto y su experiencia, o una presentación final de los resultados de minería de datos.

Salida **Informe definitivo**

En el final del proyecto, habrá al menos un informe definitivo en el que todos los hilos son encontrados. Así como la identificación de los resultados obtenidos, el informe también debería describir el proceso, mostrar los costos que se han encontrados, definir cualquier desviación del plan original, describir proyectos de implementación, y hacer cualquier recomendación para el futuro trabajo. El contenido real detallado del informe depende muchísimo de la audiencia planeada.

Actividades

- Identificar cuales informes son necesarios (presentación de diapositiva, conclusiones de administración, detalles encontrados, explicación de los modelos, etc.)
- Analizar que tan bien se han encontrado los objetivos de minería de datos iniciales
- Identificar grupos de objetivos para el informe
- Describir en forma general las estructuras y el contenido de informe(s)
- Seleccionar conclusiones para ser incluidas en los informes
- Escribir un informe

Salida **Presentación final**

Así como un informe definitivo, puede ser necesario hacer una presentación final para concluir el proyecto- tal vez al patrocinador de dirección, por ejemplo. La presentación normalmente contiene un subconjunto del contenido de la información en el informe definitivo, estructurado de un modo diferente.

Actividades

- Decidir el grupo objetivo para la presentación final y determinar si ellos ya habrán recibido el informe definitivo
- Seleccionar cuales de los artículos del informe definitivo deberían ser incluidos en la presentación final

6.4. Revisión del proyecto

Tarea **Revisar el proyecto**

Evaluar que fue lo correcto y que fue lo errado, cual fue el éxito obtenido, y que necesidades serán mejoradas.

Salida **Documentación de experiencia**

Resumir la gran experiencia ganada durante el proyecto. Por ejemplo, trampas, accesos a información incorrecta (misleading approaches), o los puntos para seleccionar las mejores técnicas de minería de datos en situaciones similares podrían ser la parte de esta documentación. En proyectos ideales, la documentación de experiencia también cubre cualquier informe que ha sido escrito por miembros individuales del proyecto durante el proyecto.

Actividades

- Entrevistar a toda la gente significativa involucrada en el proyecto y preguntarles sobre su experiencia durante el proyecto
- Si los usuarios finales trabajan en el negocio con los resultados de minería de datos, entrevistarlos: ¿Están satisfechos? ¿Cómo podría haber sido mejor realizado? ¿Necesitan de apoyo adicional?
- Resumir la realimentación y escribir la documentación de experiencia
- Analizar el proceso (las cosas que se trabajaron bien, los errores producidos, las lecciones aprendidas, etc.)
- Documentar el proceso de minería de datos específico (¿Cómo puede los resultados y la experiencia de aplicación del modelo ser realimentado en el proceso?)
- Generalizar desde los detalles para producir la experiencia útil para proyectos futuros

IV-Las salidas del CRISP-DM

Esta sección contiene las breves descripciones de los objetivos y el contenido de los informes más importantes. Aquí, enfocamos en los informes que son significativos para comunicar los resultados de una fase a la gente no involucrada en esta fase (y posiblemente no involucrada en este proyecto). Estos no son necesariamente idénticos a las salidas como lo descrito en el modelo de referencia y la guía de usuario. El objetivo de estas salidas es mas para documentar resultados mientras se está realizando el proyecto.

1. Comprensión del negocio

Los resultados de la fase de Comprensión de Negocio pueden ser resumidos en un informe. Sugerimos las secciones siguientes:

Contexto

La sección Contexto proporciona una descripción básica del contexto de proyecto. Listar cuales áreas están trabajando en el proyecto, que problemas han sido identificados, y por qué la minería de datos parece proporcionar una solución.

Objetivos de negocio y criterios de éxito

La sección de Objetivos de negocio describe los objetivos del proyecto en términos de negocio. Para cada objetivo, Los Criterios de Éxito de Negocio, esto es, describir las medidas para determinar si realmente el proyecto ha logrado en sus objetivos.

Esta sección también debería listar los objetivos que fueron considerados, pero rechazados. El razonamiento de la selección de objetivos debería ser dado.

Inventario de recursos

La sección de Inventario de Recursos apunta para identificar el personal, fuentes de datos, instalaciones técnicas, y otros recursos que pueden ser útiles en la realización del proyecto.

Requerimientos, presunciones, y restricciones

Esta sección lista los requerimientos generales para la ejecución del proyecto: tipo de resultados de proyecto, presunciones hechas sobre la naturaleza del problema y de los datos que están siendo usados, y restricciones impuestas al proyecto.

Riesgos y contingencias

Esta sección identifica los problemas que pueden ocurrir en el proyecto, describe las consecuencias, y declaran que acciones pueden ser tomadas para reducir al mínimo tales riesgos.

Terminología

La sección de Terminología permite a la gente desconocida con los problemas que están siendo dirigidos por el proyecto para hacerse más familiar con ellos.

Costos y beneficios

Esta sección describe los costos del proyecto y predice los beneficios del negocio si el proyecto es exitoso (por ejemplo, la vuelta en la Inversión). Otros beneficios menos tangibles (por ejemplo, la satisfacción del cliente) también deberían ser destacadas.

Objetivos de minería de datos y criterios de éxito

La sección de Objetivos de Minería de datos declara los resultados del proyecto que permiten el logro de los objetivos de negocio. También como el listado de los accesos probables de minería de datos, los criterios de éxito para los resultados en términos de minería de datos, también deberían ser puestos en una lista.

Plan de proyecto

Esta sección pone en una lista las etapas para ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias. Donde sea posible, esto debería hacer explícitamente las iteraciones en gran escala en el proceso por ejemplo de minería de datos - por ejemplo, las repeticiones del modelado y fases de evaluación.

Evaluación inicial de herramientas y técnicas

Esta sección da una vista inicial de que herramientas y técnicas probablemente van a ser usadas y como. Esto describe los requerimientos para las herramientas y técnicas, ponen en una lista herramientas disponibles y técnicas, y los compara a los requerimientos.

2. Comprensión de Datos

Los resultados de la fase Comprensión de Datos por lo general son documentados en varios informes. Idealmente, estos informes serían los escritos mientras se estaban realizando las respectivas tareas. Los informes describen el conjunto de datos que es explorado durante la comprensión de datos.

Para el informe definitivo, un sumario de las partes más relevantes es suficiente.

Informe de colección de datos iniciales

Este informe describe como las diferentes fuentes de datos identificadas en el inventario fueron capturadas y extraídos.

Temas para ser cubiertos:

- Contexto de datos
- Lista de fuentes de datos con amplia área de cobertura de datos requeridos por cada uno
- Para cada fuente de datos, método de adquisición o extracción
- Problemas encontrados en adquisición de datos o extracción

Informe de descripción de datos

Cada conjunto de datos adquirido es descrito en este informe.

Temas para ser cubiertos:

- Cada fuente de datos descrita detalladamente
- Lista de tablas (puede ser sólo uno) u otros objetos de base de datos
- Descripción de cada campo, incluyendo unidades, códigos usados, etc.

Informe de exploración de datos

- Este informe describe la exploración de datos y sus resultados.
    - Temas para ser cubiertos:
    - Contexto, incluyendo los amplios objetivos de exploración de datos. Para cada área de exploración emprendida:
    - Las regularidades esperadas o patrones
    - Método de detección
    - Regularidades o patrones encontrados, esperados e inesperados
    - Cualquier otra sorpresa
    - Conclusiones para transformación de datos, limpieza de datos, y cualquier otro proceso previo
    - Conclusiones relacionadas con datos que extraen objetivos u objetivos de negocio
    - Sumario de conclusiones

Informe de calidad de datos

Este informe describe lo completo y la exactitud de los datos.

Temas para ser cubiertos:

- Contexto, incluyendo amplias expectativas sobre calidad de datos. Para cada conjunto de datos:
    - Acercar tomas para evaluar la calidad de datos
    - Los resultados de evaluación de calidad de datos
    - Sumario de conclusiones de calidad de datos

3. Preparación de Datos

Los informes en la fase de preparación de datos se enfocan en los pasos de pre-proceso que producen los datos para ser minados.

Informe de descripción de conjunto de datos

Este informe proporciona una descripción del conjunto de datos (después del pre-proceso) y el proceso por el que fue producido.

Temas para ser cubiertos:

- Contexto, incluyendo objetivos amplios y plan para el pre-proceso
- Razonamiento para inclusión/exclusión de conjunto de datos. Para cada conjunto de datos incluir:
    - La descripción del pre-proceso, incluyendo las acciones que fueron necesarias para dirigir cualquier cuestión de calidad de datos

- Descripción detallada del conjunto de datos resultante, tabla por tabla y campo por campo
- Razonamiento para inclusión/exclusión de atributos
- Descubrimientos de hechos durante el pre-proceso, y cualquier implicación para futuros trabajos
- Sumario y conclusiones

4. Modelado

Las salidas producidas durante la fase Modelado pueden ser combinadas en un informe. Sugerimos las secciones siguientes:

Modelado de presunciones

Esta sección define cualquier presunción explícita hecha sobre los datos y cualquier presunción que está implícita en la técnica de modelado a ser usado.

Diseño de prueba

Esta sección describe como los modelos son construidos, probados, y evaluados.

Temas para ser cubiertos:

- Contexto de fondo la ocupación del modelo y su relación a los objetivos de minería de datos. Para cada tarea de modelado:
  - Ampliación de la descripción del tipo de modelo y los datos que se entrenan para ser usado
  - La explicación de como el modelo será probado o evaluado
  - Descripción de cualquier dato requerido para las pruebas
  - Plan para producción de los datos de prueba si hay
  - Descripción de cualquier examen planeado de modelos por expertos en dominio o de datos
  - Sumario de plan de prueba

Descripción del modelo

Este informe describe los modelos entregados y las descripciones del proceso por el que ellos fueron producidos.

Temas para ser cubiertos:

- Descripción de modelos producidos. Para cada modelo:
  - Tipo de modelo y la relación a los objetivos de minería de datos.
  - Los parámetros de ajustes usados producir el modelo
  - Descripción detallada del modelo y cualquier rasgo especial. Por ejemplo:
- Para modelos basados por regla, listar las reglas producidas más cualquier evaluación de precisión por-regla o el modelo completo y el alcance
- Para modelos no transparentes, listar cualquier información técnica sobre el modelo (como la topología de red de los nervios) y algunas descripciones de comportamiento producidas por el proceso de modelado (como la precisión o la sensibilidad)
- Descripción del comportamiento del modelo e interpretación
  - Conclusiones en cuanto a los patrones en los datos (si hay). A veces el modelo revelará hechos importantes sobre los datos sin un proceso de evaluación separado (por ejemplo, que la salida o la conclusión están duplicadas en una de las entradas).
- Sumario de conclusiones

Evaluación del modelo

Esta sección describe los resultados de prueba de los modelos según el diseño de prueba.

Temas para ser cubiertos:

- Descripción de los procesos de evaluación y los resultados, incluyendo cualquier desviación del plan de prueba. Para cada modelo:
  - Evaluación detallada, incluyendo medidas como precisión e interpretación del comportamiento

- Cualquier comentario sobre los modelos por expertos en el dominio o de datos
- Evaluación resumida de modelos
- Ideas en por qué una cierta técnica de modelado y ciertos ajustes de parámetro conducen a resultados buenos/malos
- Evaluación sumaria del conjunto de modelos completos

5. Evaluación

Evaluación de los resultados de minería de datos en lo que respecta a criterios de éxito de negocio

Este informe compara los objetivos de minería de datos con los objetivos de negocio y los criterios de éxito de negocio.

Temas para ser cubiertos:

- Revisión de objetivos de negocio y criterios de éxito de negocio (que podría haberse cambiado durante y/o como consecuencia de la minería de datos). Para cada criterio de éxito de negocio:
  - Comparación detallada entre criterio de éxito y resultados de minería de datos
  - Conclusiones sobre aceptabilidad (achievability) de criterios de éxitos y conveniencia del proceso de minería de datos
- Revisión del éxito de proyecto:
  - ¿El proyecto ha alcanzado los objetivos originales de negocio?
  - ¿Objetivos allí nuevos de negocio deben ser dirigidos después en el proyecto o en nuevos proyectos?
  - Conclusiones para futuros proyectos de minería de datos

Revisión de proceso

Esta sección evalúa la eficacia del proyecto e identifica cualquier factor que podrían haber sido pasado por alto que debería ser tenido en cuenta si el proyecto es repetido.

Lista de posibles acciones

Esta sección hace recomendaciones en cuanto a los siguientes pasos en el proyecto.

6. Desarrollo

Plan de desarrollo

Este informe especifica el desarrollo de los resultados de minería de datos.

Temas para ser cubiertos:

- Resumen de los resultados desarrollados (derivado de los informes de Próximos Pasos)
- Descripción de plan de desarrollo

Supervisión y plan de mantenimiento

La supervisión y el plan de mantenimiento especifican como los resultados desarrollados deben ser mantenidos. Temas para ser cubiertos:

- Descripción de los resultados de desarrollo y la indicación de que los resultados pueden requerir la actualización (y el por qué). Para cada resultado desarrollado:
  - Descripción de como la actualización será provocada (por una normal actualización, por un acontecimiento de activación, por la ejecución de una supervisión)
  - Descripción de como la actualización será realizada
- Resumen de los procesos de actualización de los resultados

Informe definitivo

El informe definitivo es usado para resumir el proyecto y sus resultados.

Contenido:

- Resumen de la comprensión del negocio: contexto, objetivos, y criterios de éxito
- Sumario de proceso de minería de datos
- Resumen de los resultados de minería de datos

- Sumario de la evaluación de resultados
- Resumen del desarrollo y de los planes de mantenimiento
- Análisis Costo/Beneficio
- Conclusiones para el negocio
- Conclusiones para futura minería de datos

7. Resumen de dependencias

La siguiente tabla resume las entradas principales para los operadores. Esto no significa que solo las listas de entradas puestas deberían ser consideradas -por ejemplo, los objetivos de negocio deberían ser distribuidos a todo los operadores. Sin embargo, el operador debería dirigir cuestiones específicas elevadas por sus entradas.

Phase	Deliverable	Refers To	Closely Related To
Business Understanding	Background		
	Business Objectives	Background	Terminology
	Business Success Criteria	Business Objectives	
	Inventory of Resources		
	Requirements, Assumptions & Constraints	Business Objectives	
	Risks & Contingencies	Business Objectives; Business Success Criteria	
	Terminology	Background	Business Objectives
	Costs & Benefits	Business Objectives	Project Plan
	Data Mining Goals	Business Objectives; Requirements, Assumptions & Constraints	
	Data Mining Success Criteria	Business Success Criteria; Requirements; Assumptions & Constraints; Data Mining Goals	
	Project Plan	Business Objectives; Inventory of Resources; Requirements; Assumptions & Constraints; Risks & Contingencies	Costs & Benefits
Data Understanding	Initial Data Collection Report	Business Goals; Inventory of Resources; Data Mining Goals	
	Data Description Report	Business Goals; Initial Data Collection Report	Data Quality Report
	Data Quality Report	Business Goals; Initial Data Collection Report	Data Description Report
	Exploratory Analysis Report	Business Goals; Initial Data Collection Report	
Data Preparation	Dataset & Dataset Description	Business Goals; Data Mining Goals & Data Description Report; Data Quality Report; Exploratory Analysis Report	
Modeling	Test Design	Data Mining Goals; Data Mining Success Criteria	
	Models	Data Mining Goals	Parameter Settings
	Parameter Settings	Data Mining Goals	Models
	Model Description	Models; Parameter Settings; Test Design	
	Assessment	Data Mining Success Criteria; Test Design; Models	
Evaluation	Assessment w.r.t. Business Success Criteria	Business Success Criteria; Terminology	
	Review of Process	Business Goals; Assessment w.r.t. Business Success Criteria	
	Next Steps	Project Plan; Assessment w.r.t. Business Success Criteria	
Deployment	Deployment Plan	Business Goals; Requirements, Assumptions & Constraints	Maintenance Plan
	Maintenance Plan	Business Goals; Requirements, Assumptions & Constraints	Deployment Plan
	Final Report & Presentation	Business Goals; Terminology; Assessment w.r.t. Business Success Criteria	
	Experience Documentation	Project Plan; Review of Process	

V-Apéndice

1. Glosario/Terminología

Actividad – Es parte de una tarea en la Guía de Usuario; describe las acciones para realizar una tarea

Metodología de CRISP-DM - El término general para todos los conceptos desarrollados y definidos en el CRISP-DM

Contexto de minería de datos - Un conjunto de restricciones y presunciones, tales como el tipo de problema, las técnicas o herramientas, el dominio de aplicación

Tipos de problemas de minería de datos - Una clase de típicos problemas de minería de datos, tales como la descripción de datos y el resumen, la segmentación, las descripciones de conceptos, la clasificación, la predicción, el análisis de dependencia

Genérico - Una tarea que mantiene un cruce con todos los proyectos de minería de datos posibles

Modelo - La capacidad de aplicar algoritmos a un conjunto de datos para predecir atributos objetivos; ejecutable

Salida - El resultado tangible de la ejecución de una tarea

Fase - Un término para la parte de alto nivel del modelo de proceso CRISP-DM; consiste en tareas relacionadas

Caso del proceso - Un proyecto específico descrito en términos del modelo de proceso

Modelo de proceso - Define la estructura de proyectos de minería de datos y proporciona la guía para su ejecución; consiste en el modelo de referencia y en la guía de usuario

Modelo de referencia - Descomposición de proyectos de minería de datos en fases, tareas, y salidas

Especializado - Una tarea que hace presunciones específicas en contextos específicos de minería de datos

Tarea - Una serie de actividades para producir una o más salidas; parte de una fase

Guía de usuario - Asesoramiento específico sobre como realizar proyectos de minería de datos

2. Tipos de problemas de minería de datos

Por lo general, los proyectos de minería de datos implican una combinación de diferentes tipos de problema, que juntos solucionan el problema de negocio.

2.1. Descripción de datos y resumen

La descripción y el resumen de datos apuntan a la descripción concisa de las características de los datos, típicamente en forma elemental y agregada. Esto da al usuario una descripción de la estructura de los datos. A veces, una descripción y resumen de los datos solo puede ser un objetivo de un proyecto de minería de datos. Por ejemplo, un minorista podría estar interesado en el volumen de ventas de todas las salidas separado por categorías. Los cambios y diferencias de un período anterior podrían ser resumidos y destacados. Esta clase de problema estaría en lo mas bajo de la escala de problemas de minería de datos.

En casi todos los proyectos de minería de datos, sin embargo, la descripción y resumen de los datos son un objetivo subordinado en el proceso, típicamente en sus tempranas etapas. En el principio de un proceso de minería de datos, el usuario a menudo no conoce, ni el objetivo preciso del análisis, ni la naturaleza exacta de los datos. La exploración inicial del análisis de datos puede ayudar a los usuarios a entender la naturaleza de los datos y formar hipótesis potenciales de la información oculta. La estadística descriptiva simple y las técnicas de visualización proporcionan las primeras ideas sobre los datos. Por ejemplo, la distribución de clientes por edad y regiones geográficas sugiere que partes de un grupo de clientes necesita para ser dirigida para futuras estrategias de comercialización (marketing).

La descripción y el resumen de datos típicamente ocurren en combinación con otros tipos de problemas de minería de datos. Por ejemplo, la descripción de datos puede conducir a la postulación (presunción) de segmentos interesantes en los datos. Una vez que los segmentos son identificados y definidos, una descripción y un resumen de estos segmentos son útiles. Es aconsejable llevar a cabo una descripción y resumen de datos antes de que cualquier otro tipo de problema de minería de dato sea especificado (dirigido). En este documento, esto esta reflejado en el hecho que la descripción y resumen de datos es una tarea en la fase de comprensión de datos.

El resumen también juega un papel importante en la presentación de los resultados finales. Los resultados de otros tipos de problemas de minería de datos (por ejemplo, las descripciones de conceptos o los modelos de predicción) también pueden ser considerados resumen de datos, pero sobre un nivel conceptual más alto.

Muchos sistemas de informe, paquetes estadísticos, OLAP, y sistemas EIS pueden cubrir la descripción y resumen de datos, pero hacerlo usualmente no proporciona algunos métodos para realizar modelado más avanzado. Si la descripción y resumen de datos son considerados un tipo de problema independiente y ningún modelado futuro es requerido, entonces estas herramientas pueden ser apropiadas para realizar los compromisos de minería de datos.

2.2. Segmentación

La segmentación apunta a la separación de los datos en subgrupos o clase significativos e interesantes. Todos los miembros de un subgrupo comparten características comunes. Por ejemplo, en el análisis de cesta de compras, uno podría definir los segmentos de cestas según los artículos que ellos contienen.

La segmentación puede ser realizada a mano o semi-automáticamente. El analista puede suponer ciertos subgrupos como relevantes para la pregunta de negocio, basada sobre un conocimiento previo o sobre el resultado de la descripción y el resumen de datos. En adición, hay también técnicas automáticas de agrupamiento (clustering) que pueden descubrir las estructuras antes insospechadas y ocultas en datos que permite la segmentación.

La segmentación a veces puede ser un objetivo de minería de datos. Entonces la detección de segmentos sería el objetivo principal de un proyecto de minería de datos. Por ejemplo, todas las direcciones en áreas de código postal con

la edad mas alta que el promedio y un ingreso podrían ser seleccionadas para enviar publicidad para seguro de clínica de ancianos.

Muy a menudo, sin embargo, la segmentación es un paso hacia la solución de otros tipos de problema. Entonces, el objetivo es de guardar (mantener) el tamaño de los datos manejables o encontrar los subconjuntos de datos homogéneos que son más fáciles para analizar. Típicamente en grandes conjuntos de datos variados afectan el alcance de cada uno y obscurece los patrones interesantes. Entonces, la segmentación apropiada hace la tarea más fácil. Por ejemplo, analizar las dependencias entre artículos en millones de cestas de compras es muy difícil. Esto es mucho más fácil (y más significativo, generalmente) para identificar dependencias en los segmentos interesantes de cestas de compras -por ejemplo, cestas de alto valor, cestas que contienen bienes de confort, o cestas de un día o de un periodo particular.

Nota: En la literatura, hay algo de ambigüedad en el significado de ciertos términos. A veces llaman a la segmentación agrupamiento (clustering) o clasificación (classification). El último término es confuso porque algunas personas lo usan para referirse a la creación de clases, mientras que otros piensan en la creación de modelos para predecir las clases conocidas para casos antes no vistos. En este documento, restringimos el término de clasificación al último significado (vea abajo) y usar el término segmentación con el antiguo significado, aunque las técnicas de clasificación puedan ser usadas para obtener descripciones de los segmentos descubiertos.

Técnicas apropiadas:

- Técnicas de agrupamiento (clustering)
- Redes Neuronales
- Visualización

Ejemplo:

Una empresa de venta de autos con regularidad recoge información sobre sus clientes acerca de sus características socioeconómicas como el ingreso, la edad, el sexo, la profesión, etc. Usando análisis de agrupamiento, la empresa puede dividir a sus clientes en subgrupos más comprensibles y analizar la estructura de cada subgrupo. Estrategias de control de comercialización (marketing) específicas son desarrolladas para cada grupo separado.

2.3. Descripciones de concepto

La descripción de concepto apunta a una descripción comprensible de conceptos o clases. El objetivo no es para completar el desarrollo de modelos con predicción de exactitud alta, sino para ganar ideas. Por ejemplo, una empresa puede estar interesada en el estudio sobre sus clientes más leales y desleales. De una descripción de concepto de estos conceptos (clientes leales y desleales) la compañía infiere que podría estar hecho para encontrar clientes leales o transformar clientes desleales a clientes leales.

Una descripción de concepto tiene una conexión cercana tanto a la segmentación como a la clasificación. La segmentación puede conducir a una enumeración de objetos que pertenecen a un concepto o clase sin proporcionar cualquier descripción comprensible. Típicamente la segmentación es llevada a cabo antes de que la descripción de concepto sea realizada. Algunas técnicas -técnicas de agrupamiento conceptuales, por ejemplo -ejecutan la segmentación y descripción de concepto al mismo tiempo.

Las descripciones de concepto también pueden ser usadas para objetivos de clasificación. Por otra parte, algunas técnicas de clasificación producen modelos de clasificación comprensibles, que pueden entonces ser consideradas descripciones de concepto. La distinción importante es que la clasificación apunta a ser completa en algún sentido. El modelo de clasificación tiene que aplicarse a todos los casos en la población seleccionada.

De otra manera, las descripciones de concepto no tienen que ser completas. Es suficiente si ellos describen las partes importantes de los conceptos o clases. En el ejemplo mencionado, puede ser suficiente conseguir las descripciones de conceptos de aquellos clientes que son claramente leales.

Técnicas apropiadas:

- Métodos de inducción de reglas
- Agrupamiento conceptual

Ejemplo:

Usando datos sobre los compradores de coches nuevos y una técnica de inducción de regla, una empresa de coche podría generar las reglas que describen sus clientes leales y desleales. Debajo son los ejemplos de las reglas generadas:

Si SEXO = macho y EDAD > 51 entonces CLIENTE = leal

Si SEXO = hembra y EDAD > 21 entonces CLIENTE = leal

Si PROFESIÓN = gerente y EDAD < 51 entonces CLIENTE = desleal

Si ESTADO CIVIL = soltero y EDAD < 51 entonces CLIENTE = desleal

2.4. Clasificación

La clasificación asume que hay un conjunto de objetos caracterizados por algún atributo o rasgo que pertenece a diferentes clases. La etiqueta de clase es un valor (simbólico) discreto y es conocido para cada objeto. El objetivo es para construir los modelos de clasificación (a veces llamados clasificadores), que asigna la etiqueta de clase correcta a objetos antes no vistos y sin etiquetas.

Los modelos de clasificación sobre todo son usados para el modelado predictivo.

Las etiquetas de clase pueden ser presentadas en el avance -definida por el usuario, por ejemplo, o derivadas de la segmentación. La clasificación es uno de los tipos de problemas más importantes de minería de datos que ocurren en una amplia gama de aplicaciones. Muchos problemas de minería de datos pueden ser transformados a problemas de clasificación. Por ejemplo, intentando guardar créditos para evaluar el riesgo de acreditar a un cliente nuevo. Esto puede ser transformado a un problema de clasificación para crear dos clases, clientes buenos y clientes malos. Un modelo de clasificación puede ser generado de los datos de cliente existentes de acuerdo a su comportamiento crediticio. Este modelo de clasificación puede entonces ser usado para asignar a clientes nuevos a una de las dos clases y aceptarlo o rechazarlo.

La clasificación tiene conexiones a casi todos los otros tipos de problemas. Los problemas de predicción pueden ser transformados a los problemas de clasificación por discretización de etiquetas de clase continuas, porque las técnicas de discretización permiten transformar rangos continuos en intervalos discretos. Estos intervalos discretos, más que los valores numéricos exactos, son usados como etiquetas de clase, y de ahí conducen a un problema de clasificación. Algunas técnicas de clasificación producen una clase comprensible o descripciones de concepto. Hay también una conexión al análisis de dependencia porque los modelos de clasificación típicamente usan (explotan) y aclaran las dependencias entre atributos.

La segmentación puede también proporcionar las etiquetas de clase o restringir el conjunto de datos para que buenos modelos de clasificación puedan ser construidos. Es útil analizar desviaciones antes de que un modelo de clasificación sea construido. Las desviaciones y contingencias (cosas fuera de lugar-outliers) pueden obscurecer el patrón que podría permitir un buen modelo de clasificación. De otro modo, un modelo de clasificación también puede ser usado para identificar desviaciones y otros problemas con los datos.

Técnicas apropiadas:

- Análisis de discriminante
- Métodos de inducción de regla
- Aprendizaje por árboles de Decisión
- Redes neuronales
- La K más cercana
- Razonamiento basado en caso
- Algoritmos genéticos

Ejemplo:

Los bancos generalmente tienen información sobre el comportamiento de pago de sus aspirantes de crédito. Combinando esta información financiera con otra información sobre los clientes, como el sexo, la edad, el ingreso, etc., es posible desarrollar un sistema para clasificar a clientes nuevos como clientes buenos o malos (esto es, el riesgo de crédito en la aceptación de un cliente es alto o bajo).

2.5. Predicción

Otro tipo de problema importante que ocurre en una amplia gama de usos es la predicción. La predicción es muy similar a la clasificación.

La única diferencia es que en la predicción el atributo objetivo (la clase) no es un atributo cualitativo discreto, pero es uno continuo.

El objetivo de la predicción esta en encontrar el valor numérico del atributo objetivo para objetos no vistos. En la literatura, este tipo de problema es a veces llamado regresión. Si la predicción trata con datos de serie tiempo, entonces a menudo lo llaman pronosticación.

Técnicas apropiadas:

- Análisis de regresión
- Árboles de regresión
- Redes neuronales

- La K más cercana
- Métodos de la Caja-Jenkins
- Algoritmos genéticos

Ejemplo:

El rédito anual de una empresa internacional esta correlacionado con otros atributos como la promoción, la tasa de cambio, la tasa de inflación, etc. Teniendo estos valores (o estimaciones confiables), la empresa puede predecir su rédito esperado durante el próximo año.

2.6. Análisis de dependencia

El análisis de dependencia consiste en encontrar un modelo que describe dependencias significativas (o asociaciones) entre artículos de datos o acontecimientos. Las dependencias pueden ser usadas para predecir el valor de unos datos de artículo dada la información sobre otros artículos de datos. Aunque las dependencias pueden ser usadas para el modelado predictivo, aquellos son mas usados por su comprensión. Las dependencias pueden ser estrictas o probabilísticas.

Las asociaciones son un caso especial de dependencias, que recientemente se han hecho muy populares. Las asociaciones describen las afinidades de artículos de datos (esto es, artículos de datos o los acontecimientos que con frecuencia ocurren juntos). Un típico escenario de aplicación para asociaciones es el análisis de cestas que hacen compras. Allí, una regla como “en el 30 por ciento de todas las compras, la cerveza y cacahuetes han sido comprados juntos” es un ejemplo típico para una asociación.

Los algoritmos para detectar asociaciones son muy rápidos y producen muchas asociaciones. Seleccionar el más interesante es un desafío.

El análisis de dependencia tiene conexiones cercanas a la predicción y a la clasificación, ya que las dependencias implícitamente son usadas para la formulación de modelos predictivos. Hay también una conexión a descripciones de concepto, que a menudo destacan dependencias.

En aplicaciones, el análisis de dependencia a menudo co-ocurre con la segmentación. En grandes conjunto de datos, las dependencias son raras veces significativas porque muchas influencias cubren el uno al otro. En tales casos, es aconsejable realizar un análisis de dependencia sobre más segmentos homogéneos de datos.

El modelo secuencial es una clase especial de dependencia en las que el orden de acontecimientos es considerado. En un análisis de cesta de compras, las asociaciones describen dependencias entre artículos en un tiempo dado. El patrón secuencial describe el modelo que hace compras de un cliente particular o un grupo de clientes en el tiempo.

Técnicas Apropiadas:

- Análisis de correlación
- Análisis de regresión
- Reglas de asociación
- Redes bayesianas
- Programación de lógica inductiva
- Técnicas de visualización

Ejemplo 1:

Usando el análisis de regresión, un analista de negocio ha encontrado que hay dependencias significativas entre las ventas totales de un producto y tanto en su precio como en la cantidad de gastos de publicidad. Este conocimiento permite al negocio alcanzar el nivel deseado de las ventas por cambio del precio del producto y/o el gasto de publicidad.

Ejemplo 2:

Aplicando algoritmos de regla de asociación a datos sobre accesorios de coche, una empresa de coches ha encontrado que en el 95 por ciento de casos, si un CD player es ordenado, una transmisión automática es ordenada también. Basado en esta dependencia, la empresa de coche decide ofrecer estos accesorios como un paquete, que conduce a la reducción del costo.