1. The Problem with Data Science

Operaciones de datos ⋮ 53-68 minutes
DOI: 10.1007/978-1-4842-5104-1_1, Show Details

1. El problema de la ciencia de datos

prácticas de Harvinder Atwal https://doi.org/10.1007/978-1-4842-5104-1_1

1. El problema de la ciencia de datos

Harvinder Atwal [1]

(1)

Isleworth, Reino Unido

Antes de adoptar DataOps como solución, es importante comprender el problema que estamos tratando de resolver. Cuando ve artículos en línea, escucha presentaciones en conferencias o lee sobre el éxito de organizaciones líderes impulsadas por datos como Facebook, Amazon, Netflix y Google (FANG) , entregar ciencia de datos exitosa parece un proceso simple. La realidad es muy diferente.

Si bien es indudable que existen historias de éxito, también hay muchas pruebas de que una inversión sustancial en ciencia de datos no está generando los rendimientos esperados para la mayoría de las organizaciones. Hay múltiples causas, pero se derivan de dos causas fundamentales. Primero, un enfoque de arquitectura de la información del siglo XX para el manejo de datos y análisis en el siglo XXI. En segundo lugar, la falta de conocimiento y apoyo organizacional para la ciencia y el análisis de datos. Los mantras comunes (del siglo XX) adoptados en la industria para superar estos problemas empeoran las cosas, no mejoran.

¿Hay algún problema?

Es posible crear una ventaja competitiva y resolver problemas valiosos utilizando datos. Muchas organizaciones están logrando generar historias de éxito legítimas a partir de sus inversiones en ciencia de datos y análisis de datos:

- El vicepresidente de innovación de productos, Carlos Uribe-Gómez, y el director de productos, Neil Hunt, publicaron un documento que dice que algunos de sus

algoritmos de recomendación le ahorran a Netflix mil millones de dólares cada año en una reducción de la rotación. [1]

- Una de las iniciativas de ciencia de datos de Monsanto para mejorar el transporte y la logística global ofrece ahorros anuales y reducción de costos de casi $ 14 millones, al tiempo que reduce las emisiones de C02 en 350 toneladas métricas (TM). [2]

- DeepMind de Alphabet, mejor conocido por su programa AlphaGo, ha desarrollado un sistema de inteligencia artificial (IA) en asociación con el Moorfield Eye Hospital de Londres para derivar el tratamiento de más de 50 enfermedades que amenazan la vista con tanta precisión como los médicos expertos líderes en el mundo. [3]

Al no querer quedarse atrás, la mayoría de las organizaciones ahora están gastando mucho en tecnología costosa y contratando costosos equipos de científicos de datos, ingenieros de datos y analistas de datos para dar sentido a sus datos e impulsar decisiones. Lo que alguna vez fue una actividad de nicho incluso en las organizaciones más grandes, ahora se considera una competencia central. Las tasas de crecimiento de la inversión y la posición laboral son asombrosas considerando que el PIB mundial solo está creciendo a un 3,5% anual:

- International Data Corp. (IDC) espera que los ingresos mundiales por big data y soluciones de análisis empresarial alcancen los $ 260 mil millones en 2022, una tasa de crecimiento anual compuesta del 11,9% durante el período 2017-2022. [4]

- Los informes de trabajos emergentes de LinkedIn clasifican a los ingenieros de aprendizaje automático, científicos de datos e ingenieros de big data como tres de los cuatro trabajos de más rápido crecimiento en los Estados Unidos entre 2012 y 2017. ¡Los roles de científicos de datos aumentaron más del 650% durante ese período! [5]

La realidad

A pesar del enorme desembolso monetario, solo una minoría de organizaciones logra resultados significativos. Los estudios de casos que demuestran resultados cuantificables son excepciones aisladas, que incluso permiten la renuencia a revelar ventajas competitivas. El crecimiento exponencial en el volumen de datos, los rápidos aumentos en el gasto en soluciones y las mejoras en la tecnología y los algoritmos no han llevado a un aumento en la productividad del análisis de datos.

Hay indicios de que la tasa de éxito de los proyectos de análisis de datos está disminuyendo. En 2016, Forrester concluyó que solo el 22% de las empresas experimentaron un alto crecimiento de los ingresos y se beneficiaron de sus inversiones en ciencia de datos. [6] Además, en 2016, Gartner estimó que el 60% de los proyectos de

big data fracasan, pero empeora. En 2017, Nick Heudecker de Gartner emitió una corrección. La estimación del 60% fue "demasiado conservadora", la tasa real de fallas estuvo más cerca del 85%. [7] Aunque muchos de los datos de la encuesta están relacionados con “big data”, sigo pensando que los resultados de Nick son relevantes. Fuera del campo de la ciencia de datos, la mayoría de la gente piensa erróneamente en big data, ciencia de datos y análisis de datos como términos intercambiables y responderán como tales.

Puede haber múltiples razones para la escasa tasa de rendimiento y la imposibilidad de mejorar la productividad a pesar de una importante inversión en ciencia y análisis de datos. El crecimiento explosivo en la captura de datos puede resultar en la adquisición de datos de valor marginal cada vez más bajos. Es posible que la tecnología, las bibliotecas de software y los algoritmos no sigan el ritmo del volumen y la complejidad de los datos capturados. Los niveles de habilidad de los científicos de datos podrían ser insuficientes. Es posible que los procesos no estén evolucionando para aprovechar las oportunidades impulsadas por los datos. Por último, las barreras organizativas y culturales podrían estar impidiendo la explotación de datos.

Valor de los datos

No hay indicios de que haya disminuido el valor marginal de los datos recopilados. Gran parte de los datos adicionales capturados provienen cada vez más de nuevas fuentes, como sensores de dispositivos de Internet de las cosas (IoT) o dispositivos móviles, datos no estructurados y documentos de texto, imágenes o semiestructurados generados por registros de eventos. El mayor volumen y variedad de datos adquiridos está ampliando la oportunidad para que los científicos de datos extraigan conocimiento e impulsen decisiones.

Sin embargo, existe evidencia de que la mala calidad de los datos sigue siendo un desafío importante. En el Informe de científicos de datos de 2018 de Figure Eight, el 55% de los científicos de datos mencionaron la calidad / cantidad de los datos de entrenamiento como su mayor desafío. [8] La tasa había cambiado poco desde el informe inaugural de 2015, cuando el 52,3% de los científicos de datos citaron datos de mala calidad como su mayor obstáculo diario. [9] Los datos sucios también se citaron como la barrera número uno en la Encuesta sobre el estado de los datos y el aprendizaje automático de Kaggle de 2017 a 16,000 encuestados, mientras que "datos no disponibles o de difícil acceso" fue la quinta barrera más significativa y fue mencionada por el 30,2% de los encuestados. [10]

Tecnología, software y algoritmos

No hay indicios de que la tecnología, las bibliotecas de software y los algoritmos no estén a la altura del volumen y la complejidad de los datos capturados. Las bibliotecas de tecnología y software continúan evolucionando para manejar problemas cada vez más

desafiantes al tiempo que agregan interfaces simplificadas para ocultar la complejidad a los usuarios o aumentar la automatización. Donde antes la ejecución de un clúster de Hadoop en las instalaciones era la única opción para trabajar con varios terabytes de datos, ahora se pueden ejecutar las mismas cargas de trabajo en motores de consultas Spark o SQL administrados como servicio en la nube sin ingeniería de infraestructura. requisito.

Las bibliotecas de software como Keras facilitan mucho el trabajo con bibliotecas de aprendizaje profundo como el popular TensorFlow de Google. Proveedores como DataRobot han automatizado la producción de modelos de aprendizaje automático. Los avances en los algoritmos y arquitecturas de aprendizaje profundo y las grandes redes neuronales con muchas capas, como las redes neuronales convolucionales (CNN) y las redes de memoria a corto plazo (redes LSTM), han permitido un cambio radical en el procesamiento del lenguaje natural (NLP), traducción automática, reconocimiento de imágenes, procesamiento de voz y análisis de video en tiempo real . En teoría, todos estos desarrollos deberían mejorar la productividad y el retorno de la inversión.(ROI) de la inversión en ciencia de datos. Quizás las organizaciones estén utilizando tecnología obsoleta o incorrecta.

Científicos de datos

Como campo relativamente nuevo, la inexperiencia de los científicos de datos puede ser un problema. En la Encuesta sobre el estado de los datos y el aprendizaje automático de Kaggle, el rango de edad modal de los científicos de datos era solo de 24 a 26 años, y la edad media era de 30. La edad media variaba según el país; para los Estados Unidos, fue 32. Sin embargo, esto es todavía mucho más bajo que la edad promedio del trabajador estadounidense a los 41 años. Sin embargo, el nivel educativo no fue un problema, el 15,6% tenía un doctorado, el 42% tenía una maestría y el 32% una licenciatura. [10] Dado que todas las formas de análisis avanzado eran marginales antes de 2010, también hay una deficiencia de gerentes experimentados. Como resultado, tenemos muchos científicos de datos extremadamente brillantes sin experiencia en el manejo de la cultura organizacional. y falta de liderazgo analítico superior.

Procesos de ciencia de datos

Es un desafío encontrar datos de encuestas sobre los procesos y metodologías utilizados para entregar ciencia de datos. La encuesta de KDnuggets de 2014 mostró que el proceso estándar de la industria para la minería de datos (CRISP-DM) es la principal metodología para proyectos de análisis, minería de datos y ciencia de datos utilizada por el 43% de los encuestados. [11] El siguiente enfoque más popular no fue un método en absoluto, sino los encuestados que siguieron su proceso de cosecha propia. El modelo Sample, Explore, Modify, Model and Assess (SEMMA) del Instituto SAS ocupó el tercer lugar, pero en rápido declive ya que el uso está estrechamente vinculado a los productos SAS.

El desafío con CRISP-DM y otras metodologías de minería de datos como Knowledge Discovery Databases (KDD) es que tratan la ciencia de datos como un proceso mucho más lineal de lo que es. Animan a los científicos de datos a dedicar mucho tiempo a planificar y analizar para una única entrega casi perfecta, que puede no ser lo que el cliente desea en última instancia. No se centra la atención en el producto mínimo viable, los comentarios de los clientes o la iteración para asegurarse de que está dedicando el tiempo sabiamente a trabajar en lo correcto. También tratan la implementación y el monitoreo como un problema de "tirar por la borda", donde el trabajo se pasa a otros equipos para su finalización con poca comunicación o colaboración, lo que reduce las posibilidades de una entrega exitosa.

En respuesta, muchos grupos han propuesto nuevas metodologías, incluido Microsoft con su Proceso de ciencia de datos en equipo (TDSP) . [12] TDSP es una mejora significativa con respecto a los enfoques anteriores y reconoce que la entrega de la ciencia de datos debe ser ágil, iterativa, estandarizada y colaborativa. Desafortunadamente, TDSP no parece estar ganando mucha tracción. TDSP y metodologías similares también están restringidas al ciclo de vida de la ciencia de datos. Existe la oportunidad de una metodología que abarque el ciclo de vida de los datos de un extremo a otro, desde la adquisición hasta la jubilación.

Cultura organizacional

Los factores emocionales, situacionales y culturales influyen en gran medida en las decisiones comerciales. La encuesta de FORTUNE Knowledge Group a más de 700 ejecutivos de alto nivel de una variedad de disciplinas en nueve industrias demuestra las barreras para la toma de decisiones basada en datos. La mayoría (61%) de los ejecutivos está de acuerdo en que, al tomar decisiones, las percepciones humanas deben preceder a los análisis rigurosos. El sesenta y dos por ciento de los encuestados afirma que a menudo es necesario confiar en los "sentimientos viscerales" y que los factores suaves deben tener el mismo peso que los factores duros. Es preocupante que dos tercios (66%) de los ejecutivos de TI afirmen que las decisiones a menudo se toman por el deseo de ajustarse a "la forma en que siempre se han hecho las cosas". [13]Estos no son hallazgos aislados. La Encuesta Ejecutiva de Big Data 2017 de NewVantage Partners encontró que los desafíos culturales siguen siendo un impedimento para la adopción empresarial exitosa:

> Más del 85% de los encuestados informan que sus empresas han iniciado programas para crear culturas basadas en datos, pero solo el 37% informa haber tenido éxito hasta el momento. La tecnología de big data no es el problema; la comprensión de la gestión, la alineación organizacional y la resistencia organizacional general son los culpables. Si las personas fueran tan maleables como los datos. [14]

No es de extrañar que muy pocas empresas hayan seguido el ejemplo de Amazon y hayan reemplazado a los responsables de la toma de decisiones de cuello blanco altamente remunerados con algoritmos a pesar del enorme éxito que ha logrado. [15]

El desafío de entregar una ciencia de datos exitosa tiene mucho menos que ver con la tecnología, sino con la actitud cultural en la que muchas organizaciones tratan la ciencia de datos alternativamente como un ejercicio de marcar casillas o como parte de la búsqueda interminable de la solución perfecta para todos sus desafíos. El problema tampoco es la eficacia de los algoritmos. Los algoritmos y la tecnología están muy por delante de nuestra capacidad para alimentarlos con datos de alta calidad, superar las barreras de las personas (habilidades, cultura y organización) e implementar procesos centrados en los datos. Sin embargo, estos síntomas son en sí mismos el resultado de causas profundas más profundas, como la falta de conocimiento de la mejor manera de usar los datos para tomar decisiones, las percepciones heredadas del enfoque del siglo pasado para manejar datos y brindar análisis, y la escasez de soporte para el análisis de datos.

La brecha del conocimiento

Las múltiples lagunas de conocimiento dificultan la integración de la ciencia de datos en las organizaciones cuando la implementación comienza en la parte superior de una organización. No obstante, es demasiado fácil culpar a los líderes empresariales y a los profesionales de TI por su incapacidad para generar resultados. La brecha de conocimiento es una vía de doble sentido porque los científicos de datos deben compartir la culpa.

La brecha de conocimiento de los científicos de datos

La ciencia de datos tiene como objetivo facilitar mejores decisiones, lo que lleva a acciones beneficiosas al extraer conocimiento de los datos. Para permitir mejores decisiones, los científicos de datos necesitan una buena comprensión del dominio empresarial.para permitir una mejor comprensión del problema comercial, identificar los datos correctos y prepararlos (a menudo detectando problemas de calidad por primera vez), emplear los algoritmos correctos en los datos, validar sus enfoques, convencer a las partes interesadas para que actúen, operacionalizar sus resultados y medir resultados. Esta amplitud de alcance requiere una amplia gama de habilidades, como la capacidad de colaborar y coordinarse con múltiples funciones dentro de la organización, además de su propia área de trabajo, pensamiento crítico y científico, habilidades de codificación y desarrollo de software, y conocimiento de una amplia gama de aprendizaje automático y algoritmos estadísticos. Además, la capacidad de comunicar ideas complejas a una audiencia no técnica y perspicacia para los negocios en un entorno comercial es crucial. En la profesión de la ciencia de datos,

Dado que encontrar un unicornio es raro, si no imposible (no se registran en LinkedIn ni asisten a reuniones), las organizaciones intentan encontrar la mejor opción. Contratan a personas con habilidades de programación (Python o R), análisis, aprendizaje automático y estadísticas y ciencias de la computación, que son las cinco habilidades más buscadas por los empleadores. [16] Estas habilidades deberían ser el diferenciador entre los científicos de datos y todos los demás. Desafortunadamente, esta creencia refuerza la convicción errónea entre los científicos de datos jóvenes de que las habilidades técnicas especializadas deberían ser el foco, pero esto tiende a crear equipos peligrosamente homogéneos.

> "Créame un modelo de atención de traducción automática utilizando una memoria bidireccional a largo plazo a corto plazo (LSTM) con una capa de atención que se envía a un LSTM posterior a la atención apilado que alimenta una capa softmax para las predicciones", dijo ningún CEO, nunca.

Siempre me sorprende la cantidad de candidatos que dicen que su objetivo principal es un rol que les permite crear modelos de aprendizaje profundo, aprendizaje reforzado y [inserte su algoritmo aquí]. Su objetivo no es resolver un problema real o ayudar a los clientes, sino aplicar la técnica más novedosa de la actualidad. La ciencia de datos es demasiado valiosa para tratarla como un pasatiempo pagado.

Existe una desconexión entre las habilidades que los científicos de datos creen que necesitan y lo que realmente necesitan. Desafortunadamente, las habilidades técnicas no son lo suficientemente cerca como para impulsar el éxito real y las acciones beneficiosas a partir de decisiones basadas en datos. Enfrentados a datos de mala calidad de difícil acceso, falta de apoyo de gestión, sin preguntas claras que responder o resultados ignorados por los responsables de la toma de decisiones, los científicos de datos sin liderazgo sénior en ciencia de datos no están equipados para cambiar la cultura. Algunos buscan pastos más verdes y consiguen un nuevo trabajo, solo para darse cuenta de que existen desafíos similares en la mayoría de las organizaciones. Otros se centran en la parte del proceso que pueden controlar, el modelado.

En ciencia de datos, hay un énfasis excesivo en el aprendizaje automáticoo aprendizaje profundo, y especialmente entre los científicos de datos jóvenes, la creencia de que trabajar en aislamiento solitario para maximizar la (s) puntuación (es) de precisión del modelo en un conjunto de datos de prueba es la definición de éxito. Este comportamiento es fomentado por cursos de capacitación, artículos en línea y especialmente Kaggle. La precisión alta de la predicción del conjunto de pruebas me parece una interpretación extraña del éxito. Por lo general, en mi experiencia, es mejor probar diez escenarios de solución en lugar de pasar semanas optimizando prematuramente la solución de un solo problema porque no sabe de antemano qué va a funcionar. Cuando recibe comentarios de los consumidores y mide los resultados, ve si ha impulsado una acción beneficiosa o incluso un aprendizaje útil. En ese momento, puede decidir el valor de realizar un mayor esfuerzo para optimizar.

El objetivo debe ser conseguir que se produzca un producto mínimo viable. Un modelo perfecto en una computadora portátil que nunca entra en producción desperdicia esfuerzo, por lo que es peor que un modelo que no existe. Hay dominios en los que la precisión del modelo es primordial, como la precisión del diagnóstico médico, la detección de fraudes y AdTech, pero estos son una minoría en comparación con las aplicaciones en las que hacer cualquier cosa es una mejora significativa sobre no hacer nada. Incluso en los dominios que se benefician desproporcionadamente de la optimización de la precisión del modelo, la cuantificación del impacto en el mundo real es aún más importante.

Poner un modelo en producción requiere habilidades técnicas diferentes a las de crear el modelo. Las más importantes son las habilidades de desarrollo de software relevantes. Para muchos científicos de datos, que principalmente no tienen experiencia en desarrollo de software, la codificación es solo un medio para un fin. No saben que la codificación y la ingeniería de software son disciplinas con su propio conjunto de mejores prácticas. Alternativamente, si lo saben, tienden a ver la escritura de código reutilizable , el control de versiones y las pruebas o la documentación como obstáculos que deben evitarse.

Las habilidades de desarrollo débiles causan dificultades, especialmente para la reproducibilidad, el desempeño y la calidad del trabajo. Las barreras para que los modelos entren en producción no son responsabilidad exclusiva de los científicos de datos. A menudo, no tienen acceso a las herramientas y sistemas necesarios para que sus modelos entren en producción y, por lo tanto, deben depender de otros equipos para facilitar la implementación. Los científicos de datos ingenuos ignoran el abismo entre el desarrollo local y la producción basada en servidores y lo tratan como un problema de arrojarlo por encima del cerco al no pensar en las implicaciones de su elección de lenguaje de programación. Esta inexperiencia provoca fricciones y fallas evitables.

Brecha de conocimiento de TI

Superficialmente, la ciencia de datos y el desarrollo de software comparten similitudes. Ambos involucran códigos, datos, bases de datos y entornos informáticos. Entonces, los científicos de datos requieren algunas habilidades de desarrollo de software. Sin embargo, hay una distinción crucial demostrada en la Figura 1-1 que muestra la diferencia entre el aprendizaje automático y la programación regular .

Figura 1-1

Diferencia entre programación regular y aprendizaje automático

En la programación regular, las reglas o la lógica se aplican a los datos de entrada para generar una salida (Salida = f (Entradas) como Z = X + Y) basada en requisitos bien entendidos. En el aprendizaje automático, los ejemplos de resultados y sus datos de entrada junto con sus propiedades individuales, conocidas como características, alimentan un algoritmo de aprendizaje automático. Un algoritmo de aprendizaje automático intenta aprender las reglasque generan resultados a partir de insumos

minimizando una función de costo a través de un proceso de capacitación que nunca logrará una precisión perfecta en datos de la vida real. Una vez entrenada adecuadamente, la programación regular puede usar reglas del modelo de aprendizaje automático para hacer predicciones para nuevos datos de entrada. La diferencia entre la programación regular y el aprendizaje automático tiene profundas implicaciones para la calidad de los datos, el acceso a los datos, las pruebas, los procesos de desarrollo e incluso los requisitos informáticos.

La basura que entra, la basura que sale se aplica a la programación regular. Sin embargo, los datos de alta calidad son esenciales para el aprendizaje automático, ya que el algoritmo depende de buenos datos para aprender las reglas. Los datos de mala calidad darán lugar a un entrenamiento y predicciones inferiores. Generalmente, más datos permiten que un algoritmo de aprendizaje automático descifre más complejidad y genere predicciones más precisas. Además, más características inherentes a los datos permiten que un algoritmo mejore la precisión predictiva. Los científicos de datos también pueden diseñar características adicionales a partir de datos existentes basándose en el conocimiento y la experiencia del dominio.

El entrenamiento de modelos es iterativo y computacionalmente costoso. La memoria de alta capacidad y las CPU más potentes le permiten utilizar más datos y algoritmos sofisticados. Los lenguajes y bibliotecas que se utilizan para crear modelos de aprendizaje automático están especializados para el análisis de datos, normalmente los lenguajes de programación R y Python y sus bibliotecas y paquetes, respectivamente. Sin embargo, una vez que se ha creado un modelo, los procesos de implementación son mucho más familiares para un desarrollador de software.

En la programación regular , la lógica es la parte más importante. Es decir, asegurarse de que el código sea correcto es fundamental. Los entornos de desarrollo y prueba a menudo no necesitan computadoras de alto rendimiento, y los datos de muestra con cobertura suficiente son suficientes para completar las pruebas. En la ciencia de datos, tanto los datos como el código son fundamentales. No hay una respuesta correcta para probar, solo un nivel aceptable de precisión. A menudo, se requiere un código mínimo (en comparación con la programación regular) para ajustar y validar un modelo con una alta precisión de prueba (por ejemplo, 95%). La complejidad radica en garantizar que los datos estén disponibles, se comprendan y sean correctos.

Incluso para el entrenamiento, los datos deben ser datos de producción, o el modelo no predecirá bien los datos nuevos no vistos con una distribución de datos diferente. Los datos de muestra no son útiles a menos que los científicos de datos seleccionen los datos de prueba como parte de una estrategia deliberada (por ejemplo, muestreo aleatorio, validación cruzada, muestreo estratificado) específica para el problema. Aunque estos requisitos se relacionan con el aprendizaje automático, se generalizan a todo el trabajo de los científicos de datos en general.

Las necesidades de los científicos de datos a menudo son malinterpretadas por TI, incluso por aquellos que brindan apoyo como "agradables". Con frecuencia se les pide a los científicos de datos que justifiquen por qué necesitan acceso a múltiples fuentes de datos, datos de producción completos, software específico y computadoras poderosas cuando otros "desarrolladores" no los necesitan y los analistas de informes han "extraído" datos durante años simplemente ejecutando SQL consultas sobre bases de datos relacionales. TI está frustrado porque los científicos de datos no comprenden las razones detrás de las prácticas de TI. Los científicos de datos se sienten frustrados porque no es fácil justificar el valor de lo que consideran necesidades por adelantado. Más de una vez la pregunta "¿Pero por qué necesita estos datos?" ha hecho que mi corazón se hunda.

Es raro ver procesos de TI diseñados para admitir procesos analíticos avanzados. Comienza con la forma en que se capturan los datos. Muchos desarrolladores ven la captura de nuevos elementos de datos como una carga con un costo asociado en el tiempo de planificación, análisis, diseño, implementación y mantenimiento. Por lo tanto, la mayoría de las organizaciones recopilan datos principalmente para respaldar procesos operativos como la gestión de relaciones con el cliente (CRM), la gestión financiera, la gestión de la cadena de suministro, el comercio electrónico y el marketing. Con frecuencia, estos datos residen en silos separados, cada uno con su estricta estrategia de gobierno de datos.

A menudo, los datos pasarán a través de un ETL (Extract, Transform, Load) proceso para transformarla en datos estructurados (típicamente un formato tabular) antes de cargarlo en un almacén de datos para que sea accesible para el análisis. Hay algunos inconvenientes en este enfoque para la ciencia de datos. Solo un subconjunto de datos se abre paso a través del ETL donde generalmente se prioriza para los informes. Agregar nuevos elementos de datos puede llevar meses de desarrollo. Como tal, los datos sin procesar no están disponibles para los científicos de datos. ¡Los datos brutos son lo que necesitan!

Los almacenes de datos tradicionales generalmente solo manejan datos estructurados en esquemas relacionales (con datos divididos en múltiples tablas que se pueden unir para evitar la duplicación y mejorar el rendimiento) y pueden tener dificultades para administrar la escala de datos que tenemos disponibles en la actualidad. Tampoco manejan casos de uso modernos que requieren datos no estructurados como texto o, a veces, incluso formatos de datos semiestructurados generados por máquina como JSON (notación de objetos JavaScript). Una solución es la creación de lagos de datos donde se almacenan los datos en el formato nativo de crudo y pasa por un ELT (extraer, de carga, Transform) proceso cuando sea necesario, con la transformación de ser dependiente del caso de uso.

Cuando un lago de datos no está disponible, los científicos de datos deben extraer los datos ellos mismos y combinarlos en una máquina local o trabajar con ingenieros de datos para construir tuberías en un entorno con las herramientas, los recursos

informáticos y el almacenamiento que necesitan. Las solicitudes de acceso a datos, entornos de aprovisionamiento e instalación de herramientas y software son a menudo responsabilidad de equipos separados con diferentes preocupaciones por la seguridad, los costos y la gobernanza. Como tal, los científicos de datos deben trabajar con diferentes grupos para implementar y programar sus modelos, paneles y API. Con procesos implementados incongruentes con las necesidades de la ciencia de datos, los costos son muy elevados.

Todo el ciclo de vida de los datos se divide en muchos equipos de TI, y cada uno de ellos de forma aislada toma decisiones racionales en función de sus objetivos de silos funcionales. Estos objetivos de silos no sirven a los científicos de datos. Para los científicos de datos que necesitan canalizaciones de datos desde los datos sin procesar hasta el producto de datos final, surgen desafíos importantes. Necesitan justificar sus requisitos y negociar con múltiples partes interesadas para completar una entrega. Incluso si tienen éxito, seguirán dependiendo de otros equipos para muchas tareas y estarán a merced de los retrasos y la priorización. Ninguna persona o función es responsable de todo el proceso, lo que genera retrasos, cuellos de botella y riesgo operativo.

La seguridad y la privacidad de los datos se mencionan ocasionalmente como obstáculos para evitar el acceso y el procesamiento de los datos. Existe una preocupación genuina por garantizar el cumplimiento de las regulaciones, respetar la privacidad del usuario, proteger la reputación, defender la ventaja competitiva y prevenir daños malintencionados. Sin embargo, estas preocupaciones también se pueden utilizar para tomar una ruta de aversión al riesgo y no implementar soluciones que permitan el uso seguro, legítimo y ético de los datos. Más típicamente, los problemas ocurren cuando se implementan políticas de privacidad y seguridad de datos sin realizar un análisis de costo-beneficio completo y sin comprender completamente el impacto en el análisis de datos.

Brecha de conocimiento tecnológico

Si bien la tecnología no es la única barrera para la implementación exitosa de la ciencia de datos, sigue siendo crucial para obtener las herramientas adecuadas. La Figura 1-2 muestra las capas típicas de hardware y software en un ciclo de vida de datos, desde los datos sin procesar hasta las aplicaciones comerciales útiles.

Figura 1-2

Capas típicas de hardware y software en el ciclo de vida de los datos

Muchos requisitos de software y hardware deben combinarse para crear productos de datos. Debe haber una comprensión holística de los requisitos y el equilibrio de la inversión en todas las capas del ciclo de vida. Desafortunadamente, es fácil concentrarse en una parte del rompecabezas en detrimento de otras. Las grandes empresas tienden a concentrarse en las tecnologías de big data que se utilizan para crear aplicaciones. Se

obsesionan con Kafka, Spark y Kubernetes, pero no brindan a sus científicos de datos acceso suficiente a los datos, las bibliotecas de software y las herramientas que necesitan. Es más probable que las organizaciones más pequeñas proporcionen a sus científicos de datos las herramientas de software que necesitan, pero es posible que no inviertan en tecnologías de almacenamiento y procesamiento, dejando el procesamiento analítico aislado en las computadoras portátiles.

Incluso si hacen la inversión correcta en herramientas, las organizaciones aún pueden subestimar los recursos de soporte necesarios para construir, mantener y optimizar la pila. Sin el talento suficiente en ingeniería de datos, gobernanza de datos, DevOps, administración de bases de datos, arquitectura de soluciones e ingeniería de infraestructura, es casi imposible utilizar las herramientas de manera eficiente.

Las organizaciones más grandes también tienden a cometer el error de tratar la ciencia de datos como un proyecto de tecnología de "cascada del viejo mundo", donde se necesita una infraestructura costosa o comprar herramientas de alto precio antes de poder comenzar a usar datos. Esta creencia proviene del siglo pasado, cuando su única opción para manejar datos a gran escala y aplicar análisis era comprar costosos sistemas en las instalaciones de las empresas.como Oracle, Teradata, IBM o SAS. El problema es que los proyectos de tecnología en cascada son lentos, costosos y muy propensos a fallar en las actividades de ciencia de datos. Incluso si cumplen, la tecnología avanza tan rápido que la solución final puede quedar obsoleta cuando se lance. No necesita un clúster de Hadoop empresarial multimillonario para comenzar. Solo necesita la funcionalidad suficiente para demostrar el valor antes de seguir invirtiendo en tecnología y personas.

Brecha de conocimiento de liderazgo

La mayoría de los ejecutivos de alto nivel no parecen comprender la ciencia de datos a pesar de que tienden a ser muy numerarios, o incluso a tener conocimientos de TI. Una razón podría ser que tienden a leer el último Harvard Business Review y lo tratan como un evangelio. Saben que empresas exitosas como Google, Amazon y Facebook tienen muchos datos, por lo que también deben almacenarlos. Parece que creen que contratar científicos de datos inteligentes con doctorados en astrofísica es todo lo que se necesita para hacer que la magia suceda, crear valor a partir de los datos y aumentar los flujos de dinero. El dinero fluye, pero, como hemos visto por la evidencia, en la dirección opuesta. El almacenamiento de datos tiene un costo y los científicos de datos no son baratos.

El miedo a perderse algo y, a veces, la presión de los inversores y las juntas para aparecer a la vanguardia del progreso lleva a los altos ejecutivos a saltar de un tren a otro. “Vista de cliente único”, “big data”, “Hadoop” y “Customer 360” son solo algunas de las palabras de moda que han explotado y se han desvanecido en los últimos años. La última palabra de moda en ciencia de datos es inteligencia artificial (IA) , que es la capacidad de las máquinas para realizar tareas similares a las de los humanos en términos de capacidad de aprendizaje, resolución de problemas y toma de decisiones

racionales. La compañía de investigación de inteligencia de mercado, CB Insights, ha analizado 10 años de transcripciones de llamadas de ganancias de empresas públicas. Informan que la IA superó a los "macrodatos" como la palabra de moda analítica a mediados de 2016. Desde entonces, han medido un tremendo crecimiento en las menciones de IA (se discutió alrededor de 791 veces en solo el tercer trimestre de 2017), lo que representa un aumento del 25% intertrimestral. [17]

No solo los ejecutivos de C-suite están interesados en la próxima gran novedad. Los periodistas impulsan una máquina exagerada para sus propósitos. Los proveedores renombran el mismo producto como impulsado por big data, luego impulsado por aprendizaje automático y ahora impulsado por inteligencia artificial para seguir subiendo la ola. Los consultores tienen interés en inflar los beneficios positivos de emplear nuevos enfoques y minimizar los costos reales y la complejidad de la implementación. Los buenos proveedores y consultores deben ser explícitos sobre las dificultades de trabajar en el ámbito de la ciencia de datos y proporcionar respuestas reales para resolver problemas reales.

Sin representación de datos en el C-suite, la ciencia de datos y sus técnicas deben parecer un arte oscuro. La falta de familiaridad hace que sea difícil separar la exageración de la realidad y comprender realmente los requisitos necesarios para maximizar los beneficios de los datos. La IA, por ejemplo, no es una caja negra mágica. Es solo una colección de algoritmos que una buena estrategia puede dar vida. La Figura 1-3 muestra la relación entre IA, aprendizaje automático y aprendizaje profundo.

Figura 1-3

La relación entre la IA y el aprendizaje automático

La inteligencia artificial no es nueva. Los algoritmos de IA se desarrollaron poco después de que las computadoras digitales estuvieran disponibles en la década de 1950. Siguieron siendo primitivos hasta que el aprendizaje automático (y el poder de procesamiento asociado) comenzaron a florecer en la década de 1990. El aprendizaje automático es conceptualmente un subconjunto de la IA donde los algoritmos que pueden aprender de los datos se consideran inteligentes. Sin embargo, son los rápidos avances recientes en el aprendizaje por refuerzo y el aprendizaje profundo, ambos subconjuntos del aprendizaje automático, los que han impulsado la IA a la corriente principal.

En el aprendizaje por refuerzo , un agente actúa en un entorno, recibe una recompensa (positiva, neutral o negativa) como retroalimentación y pasa al siguiente estado. El agente aprende las reglas para maximizar las recompensas acumulativas aprendiendo de la acción, la recompensa y el ciclo de estado. Un ejemplo es AlphaGo de DeepMind que venció al campeón humano del juego de Go. El aprendizaje profundo utiliza redes neuronales con muchas capas y arquitecturas especializadas para resolver problemas complejos. El aprendizaje profundo es lo que mucha gente considera hoy en día IA, que es la capacidad de los algoritmos de redes neuronales para resolver problemas de

percepción humana utilizando el procesamiento del lenguaje natural, el reconocimiento de imágenes, el procesamiento de voz y el análisis, los enfoques o las habilidades de la visión en tiempo real.

AIes real. Sus beneficios son reales. No es un simple cambio de marca de estadísticas, aprendizaje automático o redes neuronales, sino un superconjunto de enfoques. Sin embargo, una estrategia de IA por sí sola no arreglará mágicamente productos terribles, marketing ineficaz y procesos rotos que tienen problemas profundos. Tampoco es algo que pueda implementar sin las bases adecuadas en su lugar. Una organización que se esfuerza por incorporar mediciones y pronósticos estadísticos simples tendrá dificultades para hacer que el aprendizaje automático funcione, y una organización que no pueda lograr el éxito con el aprendizaje automático diario no encontrará fácil aprovechar el aprendizaje profundo. La falta de comprensión del liderazgo de la IA conduce a expectativas poco realistas de logros a corto plazo y al desconocimiento de la inversión necesaria no solo en las personas y la tecnología, sino también en el cambio de procesos y cultura.

Brecha de alfabetización de datos

La alfabetización de datos es necesaria para tomar decisiones a partir de los datos y / o confiar en las decisiones de las máquinas automatizadas. La alfabetización en datos es la capacidad de leer tablas y gráficos de datos, comprenderlos para sacar conclusiones correctas y saber cuándo es potencialmente desinformador. También incluye la capacidad de comunicar significado a otros a partir de los datos. Para ofrecer el máximo beneficio de los datos y hacer que la ciencia de datos sea exitosa, se necesita un conocimiento avanzado de los datos. La alfabetización avanzada de datos incluye conocimiento de diseño experimental, alfabetización estadística, comprensión de análisis predictivo, la capacidad de dar sentido a los "macrodatos" mediante el aprendizaje automático y la capacidad de extraer significado de datos no estructurados como texto, imágenes, audio y video.

Desafortunadamente, la alfabetización avanzada en datos es extremadamente rara en las organizaciones. La oferta de personas con habilidades avanzadas de alfabetización de datos debería aumentar, pero incluso entonces, no tengo esperanzas de que las habilidades se generalicen. El desafío no se limita a los datos, sino que se aplica a cualquier dominio técnico. Cada semana, me encuentro con ejemplos de colegas que llevan a cabo procesos manuales que con habilidades de codificación básicas o intermedias podrían automatizarse. En la mayoría de los casos, los beneficios serían al menos diez veces la mejora de la productividad de dicha automatización. Sin embargo, a pesar de que la capacidad de codificar en computadoras personales es ampliamente posible desde principios de la década de 1980 y la amplia disponibilidad de material de capacitación gratuito, la mayoría de las personas ven la codificación como demasiado complicada, no es su trabajo o incluso una amenaza. La inteligencia artificial y el aprendizaje automático entusiasman a algunas personas,

La falta de conocimientos avanzados de datos provoca problemas importantes. Los colegas con los conocimientos necesarios para ser peligrosos pueden tomar decisiones equivocadas. La paradoja de Simpson (sacar la conclusión incorrecta de la agregación de datos) y el sesgo de supervivencia (basar los resultados en aquellos que pasaron por un proceso de filtrado y no en la cohorte original) son solo dos ejemplos generalizados. Muchos escollos analíticos para los incautos pasan desapercibidos donde un buen científico de datos los detectará.

Los científicos de datos competentes se asegurarán de que el problema esté estructurado adecuadamente, aplicarán las técnicas correctas y sacarán las inferencias correctas. Imagine que realiza una prueba en su sitio web para medir el impacto de conversión de cambiar el color del mensaje "¡Solicite ahora!" botón. La mitad de los visitantes del sitio web al azar verán el botón verde existente (el campeón) y la mitad del nuevo botón rojo (el retador). Después de un día de prueba, la tasa de clics en el botón rojo es del 33% y en el botón verde del 32%.

El gerente de productodesea declarar otra optimización ganadora y desplegar el botón rojo, pero el científico de datos no está convencido. Quiere saber el volumen de visitantes para ejecutar una prueba de hipótesis estadística para asegurarse de que los resultados no se deben al azar en lugar de a la tasa de clics oa todos los visitantes. Es posible que algunos de los visitantes que vean el botón rojo hayan regresado recientemente y, por lo tanto, hayan visto previamente el botón verde. Como tal, no deben considerarse lo mismo que los visitantes que ven un botón rojo por primera vez. Los clics no suelen ser los lugares donde la empresa gana dinero porque muchos clientes hacen clic para averiguar los costos de envío. Es fácil influir en los clics, pero las ventas reales son mucho más difíciles de lograr.

Los científicos de datos piensan críticamente para no tomar los resultados al pie de la letra. En uno de los primeros ejemplos de ciencia de datos en la prensa popular, el minorista estadounidense Target pudo determinar que una adolescente estaba embarazada antes que el padre. [18] Entrenaron un modelo estadístico sobre las compras realizadas por mujeres antes de inscribirse en el registro de bebés de Target y utilizaron el modelo para enviar cupones específicos para productos para bebés durante el embarazo. Un padre se quejó con Target sobre los cupones que recibió su hija y luego se disculpó cuando descubrió que estaba embarazada. ¿Una victoria para el modelado estadístico? Nuestro científico de datos no está impresionado, y no solo por lo espeluznante de la campaña. Ella sabe que puede lograr el mismo resultado sin datos ni modelo. Envías a todo el mundocupones relacionados con el bebé. Al menos una de las beneficiarias seguramente estará embarazada y no se lo habrá contado a su padre. En cambio, quiere conocer las métricas de precisión del modelo, como la tasa de falsos positivos (cuántas personas se predijo erróneamente que estaban embarazadas cuando no lo estaban) antes de decidir si tiene éxito.

Depender únicamente de la alfabetización básica en datos tiende a resultar en un uso subóptimo de los datos. A menudo, las personas piensan que están tomando decisiones basadas en datos cuando en realidad están tomando decisiones basadas en hipótesis. Un ejemplo típico es el uso de una sola dimensión de datos, o criterios basados en reglas, para tomar decisiones con respecto a un objetivo específico (por ejemplo, los hombres entre las edades de 25 y 34 son los más propensos a presentar reclamos fraudulentos, por lo que deberíamos gastar nuestra auditoría presupuesto en ellos), que es anteponer el carro de datos al caballo objetivo. Un científico de datos entiende que encontrar clientes con más probabilidades de presentar reclamos fraudulentos es un objetivo para un problema de modelado predictivo que luego puede usarse para evaluar todos los datos disponibles para calcular una probabilidad mucho más precisa de fraude a nivel de cliente individual. Modelado predictivoy otras técnicas avanzadas son opacas para quienes no tienen conocimientos avanzados de datos, por lo que es fácil para ellos rechazar la solución de los científicos de datos por ser demasiado complicada.

Cuando estoy enfermo, no intento automedicarme. Visito a un profesional de la salud para pedir consejo y confío en su experiencia y recomendaciones basadas en el conocimiento. Confiar en expertos también es un comportamiento típico en las organizaciones. Los contadores elaboran informes financieros, los especialistas en adquisiciones llevan a cabo la negociación con los proveedores y los gerentes de producción dirigen la fabricación. Rara vez los responsables de la toma de decisiones desafían o cuestionan a los especialistas. Sin embargo, la alfabetización básica en datos está lo suficientemente extendida en la mayoría de las organizaciones que muchos responsables de la toma de decisiones piensan que "la entienden". Creen que están evaluando con precisión los datos junto con el instinto para tomar decisiones informadas y están explotando los datos por completo para tomar decisiones correctas. Desafortunadamente, sin una alfabetización avanzada en datos, tienden a equivocarse.

La ausencia de conocimientos avanzados de datos es la razón por la que la democratización de los datos tiene límites. Si bien más información es mejor que menos, mover datos de una hoja de cálculo de Excel a una herramienta visual de Business Intelligence (BI) no conducirá automáticamente a mejores decisiones, al igual que mostrarme imágenes de tomografía computarizada y resultados de análisis de sangre no me ayudará a curarme. No tengo las habilidades para diagnosticar y tomar la decisión correcta sobre medicamentos y procedimientos.

Falta de apoyo

Contratar gente inteligente y dejar que agreguen valor comercial es una receta para el fracaso. Científicos de datosestán encargados de averiguar sus objetivos y problemas para resolver, encontrar datos, obtener acceso, limpiar datos, instalar software y encontrar hardware para ejecutar sus trabajos orientados a datos. Posiblemente porque muchos en la organización ven la ciencia de datos como una versión más grande de la exploración de datos en libros de Excel, muchas empresas no se han dado cuenta de la

necesidad de cambiar y aprovechar un nuevo mundo de oportunidades basadas en datos. Históricamente, los departamentos de TI han bloqueado el acceso al software, los sistemas y los datos para hacer cumplir los estándares de seguridad y gobierno, y la mayoría continúa haciéndolo. La gestión de datos y la ciencia de datos no se excluyen mutuamente. La gestión de datos correcta hace que la ciencia de datos sea más eficaz. La gestión de datos incorrecta lo asfixia.

Educación y Cultura

Una expectativa común es cobrar a los científicos de datosser el único responsable de educar a la empresa para que se base en datos. Sin embargo, los científicos de datos tienden a ser relativamente inexpertos en los procesos comerciales en comparación con las personas con las que trabajan y, por lo tanto, pueden tener dificultades para cambiar la cultura. Puede haber éxito al influir en las personas (campeones) para facilitar las decisiones basadas en datos, pero si se van, el ciclo debe comenzar de nuevo con sus reemplazos. Convertir a los científicos de datos en profesores no es un buen uso de su tiempo y no esperamos que esto suceda en otras profesiones. La formación en analítica avanzada debe proporcionarse de forma centralizada a quienes la deseen. Como siempre será una pequeña minoría de personas, si realmente quiere convertirse en una empresa basada en datos, debe contratar como Google. Es decir, encontrar personas analíticas avanzadas que hagan buenos productos, comerciales,[19]

La creación de una cultura basada en datos debe guiarse de arriba hacia abajo con el ejemplo. Además de invertir en personas y tecnología, se debe considerar que los altos ejecutivos toman decisiones basadas en datos y exigen ver el mismo comportamiento de sus subordinados directos. Deben medir las acciones posteriores y solicitar ver las métricas publicadas. Este tipo de modelado de roles caerá en cascada en la organización. Desafortunadamente, esto rara vez ocurre. Los altos ejecutivos están en sus roles debido a su experiencia y juicio. A menudo, exigen que otros tomen decisiones basadas en datos mientras continúan tomando decisiones instintivas que les han servido bien en el pasado. Este comportamiento envía señales equivocadas sobre la importancia de los datos en la toma de decisiones.

Objetivos poco claros

Los datos, los conocimientos, las decisiones y las acciones no son sinónimos. Los datos son información en bruto. Las estadísticas son una comprensión precisa de esos datos. Esos conocimientos deberían conducir a la decisión de hacer algo diferente y resultar en una acción beneficiosa. Un problema común es tratar la ciencia de datos como una actividad de investigación pura. A los científicos de datos se les pide que examinen los datos en busca de "conocimientos interesantes" o que respondan "preguntas interesantes" sin un objetivo claro. La sabiduría convencional es que, dada la libertad, los científicos de datos encontrarán pepitas de conocimiento y las convertirán automágicamente en dinero. Hay varios problemas con este enfoque.

Insight que no está estrechamente alineado con un objetivo comercial y un caso de uso definido enfrenta barreras significativas para impulsar una decisión humana que lleve a la acción, dadas las dificultades culturales para lograr que los gerentes actúen sobre los datos y los obstáculos de TI en la producción de productos de datos. En el mejor de los casos, dicha información proporciona un informe de apoyo a la toma de decisiones ad hoc que puede o no conducir a la acción. Desesperados por demostrar su utilidad, los científicos de datos responden preguntas comerciales de valor cada vez más bajo que los equipos de inteligencia empresarial o el análisis de datos de autoservicio deberían responder. La organización ve poco impacto en los indicadores clave de rendimiento (KPI) y cuestiona el beneficio del costoso equipo de ciencia de datos.

Hay otra razón por la que el insight no debería ser el centro de atención. Permítanme compartirles una idea de cuando estaba en Dunnhumby. Los fundadores se dieron cuenta desde el principio de que si producían conocimiento del cliente basado en datos para diez millones de clientes, no se les pagaría diez veces más dinero que si producían conocimiento del cliente para un millón de clientes. Sin embargo, se les pagaría diez veces más si usaran datos para vender diez millones de cupones de comestibles específicos en lugar de un millón. Dunnhumby fue una de las primeras empresas del mundo en monetizar datos a escala industrial utilizando datos de esquemas de lealtad de Tesco y Kroger para vender medios dirigidos individualmente en lugar de producir información de datos de apoyo a la toma de decisiones para la administración como lo hicieron otras agencias en ese momento. Habían descubierto la importancia de escalar las decisiones basadas en datos.

Puede utilizar los mismos datos para múltiples propósitos. Los recursos en la nube pueden escalar masivamente y las canalizaciones de datos pueden automatizarse, pero las personas están limitadas por las horas que pueden trabajar. Las organizaciones deben buscar formas de escalar el impacto de sus científicos de datos. El problema con el tiempo dedicado a producir información de soporte de decisiones basada en datos es que no se escala. Si desea duplicar el beneficio, debe duplicar las horas dedicadas. Trate el conocimiento puro como cualquier investigación pura basada en laboratorio. Se le debe asignar un presupuesto de investigación y desarrollo (I + D) a partir de recursos de ciencia de datos con objetivos estratégicos específicos a largo plazo.

Si el insight no es el producto, sino el comienzo de un proceso, ¿por qué todavía se le presta tanta atención? La razón es simplemente un legado. Hubo muy poca automatización digital en el pasado, por lo que el análisis de datos intentó influir en la toma de decisiones humana. Los analistas de datos exploraron los datos, presentaron un documento con recomendaciones y pasaron al siguiente problema. Los resultados del análisis permanecieron en la computadora o en un informe escrito, pero ya no es así. Las oportunidades para tomar decisiones automatizadas sobre máquinas superan a las de la toma de decisiones humana. El momento decisivo llegó hace casi dos décadas cuando el algoritmo PageRank automatizado de Google demolió el directorio de la Web editado por humanos de Yahoo, que en ese momento era la página web más visitada del planeta.

Dejar que los científicos de datos lo averigüen

Un viaje típico de un científico de datos está lleno de frustraciones. Un científico de datos tiene un problema empresarial que resolver y uno de los primeros pasos es evaluar los datos disponibles en la organización. Le resulta difícil descubrir fuentes de datos útiles, ya que normalmente no están bien documentadas. Después de algunas conversaciones con expertos en la materia, identifica algunas fuentes de datos anónimas potencialmente útiles, pero descubre que no tiene permiso para verlas y completa una solicitud de acceso a datos de TI. El servicio de asistencia técnica de TI no comprende por qué alguien querría analizar los datos almacenados fuera del almacén de datos, pero después de alguna explicación, obtiene acceso.

Nuestra científica de datos trabaja con muestras de los datos a nivel local, ya que es el único entorno analítico al que tiene acceso. La falta de metadatos (datos sobre los datos) y la gobernanza de los datos hace que dedique un tiempo considerable a identificar el significado de los elementos de datos, depurar los datos de prueba y reformatear los campos de fecha y hora. Con el tiempo, construye un excelente modelo predictivo de abandono, y el propietario del producto está entusiasmado con la perspectiva de identificar con precisión a los clientes en riesgo de caducidad y está ansioso por probar estrategias de retención proactiva.

Para que el modelo sea útil, debe ejecutarse diariamente en una máquina más potente con una mínima intervención manual y generar recomendaciones automáticamente al sistema CRM. Nuestro científico de datos se da cuenta de que el modelo deberá ejecutarse en una máquina virtual local en el centro de datos, por lo que solicita recursos de TI. El servicio de asistencia de TI no comprende por qué alguien querría ejecutar una aplicación experimental no operativa en un servidor, pero después de una escalada acepta agregar la solicitud a su lista de trabajos pendientes.

Tres meses después, se aprovisiona el servidor. Emocionada, nuestra científica de datos accede a la máquina pero descubre que no puede instalar los lenguajes y bibliotecas que necesita porque los protocolos de seguridad bloquean Internet. Ella solicita la instalación del lenguaje de programación R junto con las bibliotecas y herramientas que necesita. El servicio de asistencia técnica de TI no reconoce el software y no puede entender por qué la "aplicación" no se puede escribir utilizando los lenguajes que utilizan otros desarrolladores, a saber, Java, JavaScript o C ++. Después de meses de mayor escalada e investigación, TI determina que el lenguaje de programación R y las bibliotecas solicitadas no son una amenaza para la seguridad y acepta la instalación.

Nuestro científico de datos ahora puede configurar la conectividad y el acceso seguro a los sistemas internos que necesita, pero durante la construcción del modelo se da cuenta de que necesita solicitar otra biblioteca. Otra solicitud al servicio de asistencia técnica de TI y algunas semanas después, tiene su biblioteca. Eventualmente, hay resultados para compartir, y el accionista empresarial está satisfecho con la reducción en la rotación y el

aumento en los ingresos logrados al retener a los clientes objetivo con incentivos de renovación.

Todos están de acuerdo en que hay razones para pasar del modelo de desarrollo experimental a una versión operativa producida mantenida por el equipo de ingeniería de datos. El código del modelo de desarrollo fue una mejora y un cambio del código de la computadora portátil y necesita refactorización para uso operativo. El equipo de ingeniería de datos no se siente cómodo con el código R en producción, así que decida recodificar en Python. Solicitan la instalación del lenguaje de programación Python junto con los paquetes y herramientas que necesitan. El servicio de asistencia técnica de TI no reconoce el software y no puede entender por qué la “aplicación” no está escrita en los lenguajes que usan otros desarrolladores, Java, JavaScript, C ++ o R….

Así continúa, barrera tras barrera, retraso tras retraso, lo que significa que una gran cantidad de análisis de datos permanece en las máquinas locales o emerge la TI en la sombra. El análisis de portátiles y la TI en la sombra no son deseables por dos razones. Primero, fomentan terribles prácticas de desarrollo. En segundo lugar, es menos probable que el trabajo termine en producción, donde puede afectar a los consumidores. La gente quiere hacer su trabajo, por lo que parecen soluciones que son peores que los problemas que los controles originales intentaban evitar.

Resumen

En este capítulo, encontramos los problemas que estamos tratando de resolver. Si bien existen logros importantes, también existe abundante evidencia de que muchas organizaciones no están generando el ROI que esperan de su inversión en ciencia de datos. Las causas principales son las lagunas de conocimiento (en ocasiones, abismos), enfoques obsoletos para administrar datos y producir análisis, y la falta de soporte para el análisis de datos dentro de la organización.

Los equipos de ciencia de datos se centran demasiado en los algoritmos y no en las habilidades de un extremo a otro que necesitan, como la colaboración con las partes interesadas y las mejores prácticas de desarrollo de software para producir un producto de datos y obtener comentarios. Como la ciencia de datos es un campo relativamente nuevo, TI no aprecia las diferencias con el desarrollo regular de software y la importancia de proporcionar una gran variedad, veracidad y volumen de datos. Los silos organizativos dividen la cadena de suministro de datos en los silos departamentales e introducen fricciones innecesarias para los científicos de datos que necesitan canales de datos eficientes.

Los científicos de datos deben superar las barreras por sí mismos, descubrir datos, acceder a datos, limpiar datos y negociar recursos informáticos y software sin soporte dedicado. Existe una subestimación de la variedad de herramientas, software y recursos de apoyo necesarios para la pila de ciencia de datos, lo que resulta en un uso ineficiente de la tecnología.

El liderazgo sénior no comprende la ciencia de datos que conduce a expectativas poco realistas y un punto ciego para la inversión necesaria en el proceso y el cambio de cultura. Se deja a los científicos de datos hacer que la organización se centre en los datos, mientras que los líderes sénior no logran modelar la toma de decisiones basada en datos. La falta de conocimientos avanzados de datos conduce a malas decisiones y al desconocimiento de las recomendaciones de ciencia de datos. Sin objetivos claros o preguntas que responder, los científicos de datos se centran en producir conocimientos en lugar de productos de datos escalables. En el próximo capítulo, comenzaremos a resolver algunos de estos problemas.

Notas finales

1. 1.

   Carlos A. Gomez-Uribe y Neil Hunt, “El sistema de recomendación de Netflix: algoritmos, valor comercial e innovación”, Computación avanzada para maquinaria, enero de 2016. https://dl.acm.org/citation.cfm?id=2843948

2. 2.
3. 3.
4. 4.

   Se prevé que los ingresos de las soluciones de Big Data y análisis de negocios alcanzarán los $ 260 mil millones en 2022, liderados por las industrias bancaria y manufacturera, según IDC, agosto de 2018. www.idc.com/getdoc.jsp?containerId=prUS44215218

5. 5.
6. 6.
7. 7.
8. 8.
9. 9.
10. 10.
11. 11.
12. 12.
13. 13.
14. 14.
15. 15.
16. dieciséis.
17. 17.
18. 18.
19. 19.