Todas las habilidades necesarias para convertirse en un científico de datos

En HostDime nos apasiona la tecnología, los emprendimientos disruptivos y la ciencia de datos en general. De ahí que nos parezca valioso este tipo de reflexiones que ofrecemos a continuación. Hace unos meses, el prestigioso sitio de búsqueda de empleo Glassdoor nombró a Data Scientist como el trabajo número uno entre sus 25 mejores trabajos del mundo. A través de este artículo, descubra las habilidades necesarias para ejercer esta profesión en el corazón del Big Data.

Responsable de la gestión, análisis y explotación de big data dentro de una empresa, el Data Scientist es la evolución del Data Analyst en la era de Big Data. Según el estudio de Glassdoor, el salario medio anual de un científico de datos es de 116.840 dólares. Teniendo en cuenta la extrema especialización necesaria para el ejercicio de esta profesión, las oportunidades laborales son numerosas y muy superiores a la cantidad de perfiles cualificados. Sin duda, el trabajo de Data Scientist es fascinante. Sin embargo, también es un puesto de alta responsabilidad, que requiere predisposiciones naturales y educación de alto nivel. Aquí están las habilidades esenciales para aspirar a una carrera en este campo.

¿Cómo convertirse en un científico de datos? Capacitación y

habilidades requeridas

Comprender los conceptos básicos de la ciencia de datos

Es imperativo que un científico de datos domine los fundamentos de la ciencia de datos. Muchos principiantes cometen el error de aplicar métodos de aprendizaje automático sin comprender los conceptos básicos. Esto es un error. El experto debe poder diferenciar el aprendizaje automático y el aprendizaje profundo, y distinguir la ciencia de datos del análisis empresarial y la ingeniería de datos. También debe conocer las herramientas más utilizadas. Finalmente, puede distinguir entre problemas de regresión y clasificación, así como aprendizaje supervisado y no supervisado.

Capacitación en análisis de datos

Actualmente, el 88% de los científicos de datos tiene al menos una maestría y el 46% de ellos tiene un doctorado. Esta educación escolar parece necesaria para desarrollar el nivel de conocimientos suficiente para el ejercicio de esta profesión. La mayoría de los profesionales (32%) tienen experiencia en matemáticas y estadística . El 19% ha estudiado informática y el 16% procede de escuelas de ingeniería.

Conocimiento de estadística

Es esencial que un científico de datos tenga al menos algunas nociones de cálculos estadísticos. Este conocimiento le permitirá determinar el enfoque y la técnica de análisis correctos para cada dato. La estadística es un concepto esencial para la producción de modelos de alta calidad, como la forma en que se usa la gramática para construir oraciones. Son la base del aprendizaje automático.

Idealmente, el profesional debería estar familiarizado con el concepto de estadística descriptiva, incluidas medias, medianas, varianza o desviación. Diferentes distribuciones de probabilidad, muestras o estadísticas inferenciales son algunos de los otros conceptos a dominar. Big Data Big Data y Data Science son dos conceptos que no deben confundirse , pero están estrechamente vinculados. De hecho, la ciencia de datos es la clave para manipular y explotar big data. Ahora generamos enormes volúmenes de datos todos los días, especialmente tras el auge de la web, las redes sociales y la IoT. La era del big data ha comenzado y muchas empresas están abrumadas por los datos. Un científico de datos debe poder procesar y analizar macrodatos.

Debe saber cómo utilizar las herramientas y tecnologías para hacer frente a estos volúmenes colosales que inducen nuevas limitaciones en términos de almacenamiento y procesamiento. Entre estas herramientas se encuentran Hadoop, Spark, Apache Storm, Flink y Hive.

Dominio de las herramientas de Big Data

Generalmente se requiere un conocimiento profundo de al menos una herramienta analítica como SAS o R. Para la ciencia de datos, se da preferencia principalmente a R, el lenguaje informático histórico y normativo para el análisis y la exploración de datos.

Lenguajes de programación

Los puestos de científico de datos requieren competencia en al menos un lenguaje de programación. El más utilizado es Python, pero puede ser reemplazado por R, Java, Julia, Pearl o C / C ++. En general, se prefiere Python porque es un lenguaje generalista con muchas bibliotecas dedicadas a la ciencia de datos. Por su parte, R es un lenguaje dedicado al análisis estadístico y visualización de datos. Julia reúne lo mejor de ambos mundos y es más rápida. El aumento de la potencia

informática de las computadoras es la fuente del auge del aprendizaje automático, y los lenguajes de programación nos permiten comunicarnos con estas máquinas. Si bien no es necesario que sea el mejor programador del mundo, un científico de datos debe saber cómo usarlos.

Saber analizar y manipular datos

Puede parecer obvio, pero un científico de datos debe sentirse perfectamente cómodo manipulando y analizando datos. El «Data Wrangling» consiste en manipular los datos, limpiarlos y transformarlos en un formato adecuado para el análisis. Este paso es necesario para simplificar el análisis de datos y mejorar sus resultados. El análisis de datos, por otro lado, se trata de aprender de los datos. Usamos para este propósito Excel, SQL o Pandas en Python . Este es el corazón del trabajo de un analista de datos, pero el de un científico de datos va más allá al utilizar el aprendizaje automático.

Visualización de datos

La visualización de datos consiste en presentar los resultados del análisis de datos en forma de gráficos , diagramas u otros diagramas. Esto hace que sea mucho más fácil para la audiencia interpretar los resultados. Hay muchas herramientas disponibles para realizar esta tarea. Los diferentes lenguajes de programación de Data Science como Python ofrecen diferentes librerías para la creación de gráficos avanzados. También podemos citar software especializado como Tableau.

Aprendizaje automático

El aprendizaje automático es la habilidad que realmente distingue al científico de datos del analista de datos. Se utiliza para crear modelos predictivos, basándose en datos del pasado para predecir tendencias futuras. Los diferentes algoritmos de Machine Learning como modelos de regresión lineal y logística resuelven diversos problemas. Un científico

de datos necesita conocer el código de cada uno de estos muchos algoritmos, pero lo que es más importante, cómo funcionan. De esta forma, puede elegir el modelo adecuado según los problemas a abordar. También puede configurar hiperparámetros y reducir la tasa de error de su modelo.

Aprendizaje profundo

El Deep Learning y las redes neuronales artificiales son una subcategoría de la inteligencia artificial, en la que se basan muchas innovaciones recientes como los vehículos autónomos. El auge de esta rama de la IA está relacionado con los avances recientes en términos de capacidad de almacenamiento y computación. Un científico de datos moderno debe tener algún conocimiento en esta área. Para dominar el Deep Learning es necesario utilizar un lenguaje de programación como Python y tener conocimientos de álgebra y matemáticas. Las bibliotecas como TensorFlow, Keras y PyTorch también son herramientas esenciales. Comprender el álgebra lineal y las funciones de varias variables. El álgebra lineal y las funciones de varias variables forman la base de muchas técnicas estadísticas computacionales y de aprendizaje automático.

Incluso si se implementa con R o sklearn, algunas empresas cuyo producto se basa en datos pueden decidir desarrollar sus propias implementaciones para mejorar sus algoritmos o rendimiento predictivo.

El uso de Hadoop

Si bien algunas empresas no lo requieren, la mayoría de las veces se requiere el dominio de la plataforma Hadoop . Asimismo, la experiencia con las herramientas de procesamiento de Hive y Pig es un argumento adicional para el reclutamiento. Las herramientas en la nube como Amazon S3 también son importantes.

Programación en SQL

Las bases de datos Hadoop y NoSQL se han establecido ampliamente en el campo de Big Data. Sin embargo, la mayoría de los reclutadores requieren que los candidatos tengan competencia en programación en SQL para poder formular y ejecutar consultas.

Gestionar datos no estructurados

Para convertirse en un Data Scientist, es fundamental saber cómo gestionar los datos no estructurados de las redes sociales, o incluso las transmisiones de vídeo o audio. Estos datos son el principal desafío del Big Data. También es importante saber cómo tratar los datos que tienen imperfecciones, como valores perdidos o cadenas de formato inconsistentes. Esta habilidad es particularmente importante en empresas que no están acostumbradas al análisis de datos.

Ingeniería de software

En una pequeña empresa que no esté familiarizada con la ciencia de datos, un científico de datos debe tener habilidades de ingeniería de software. Estos le permitirán hacerse cargo del desarrollo de un producto impulsado por datos o registro de datos. Las habilidades de ingeniería de software son esenciales para que los científicos de datos creen modelos de aprendizaje automático. El profesional debe conocer los fundamentos de la Ingeniería de Software como ciclo de vida de un proyecto de desarrollo. Saber escribir código limpio y eficiente es muy útil, y también te permite colaborar mejor con los desarrolladores y el resto de equipos de la empresa. Una base sólida es un activo valioso.

Despliegue del modelo

A menudo se pasa por alto, la implementación de modelos es un paso crucial en el aprendizaje automático. Su objetivo es

permitir que los usuarios finales utilicen el modelo, sin tener las habilidades técnicas de Data Scientist. En general, esta tarea de despliegue y puesta en producción de modelos la asume el Ingeniero de Machine Learning, lo que puede percibirse como una evolución o una especialización del Machine Learning. El científico de datos capaz de implementar modelos de aprendizaje automático aporta un valor inmenso a su negocio.

Curiosidad intelectual

La curiosidad intelectual es fundamental para detectar los datos más interesantes y explotables dentro de un volumen gigantesco de datos. Para hacer el trabajo de científico de datos con éxito, debe ser creativo y hacer sus propias preguntas en lugar de simplemente responder a las que surjan. El científico de datos debe cuestionar las causas de un evento y cómo ocurre. Debe preguntarse sobre las posibles consecuencias de cada cambio. El interrogatorio perpetuo es la habilidad blanda más importante de un científico de datos. Es esta curiosidad la que le permitirá lograr el objetivo final del proyecto de Machine Learning, y justificar los resultados de su trabajo. También le permitirá estar al tanto de las novedades en el campo de la Ciencia de Datos y seguir aprendiendo día a día.

La narración

Las tablas de datos brutos no le dicen nada a nadie. Para transmitir y compartir los resultados de sus análisis de datos, un científico de datos debe poder contar una historia en forma de visualización de datos. Los diagramas y gráficos son presentaciones interactivas que el cerebro humano puede entender de forma natural e intuitiva. La narración es una de las principales cualidades de un científico de datos.

Pensamiento estructurado

Los mejores científicos de datos pueden dividir un problema en varias partes para resolverlo de manera más efectiva. A esto se le llama pensamiento estructurado. Ésta es una cualidad muy importante para abordar problemas desde diferentes ángulos. Algunas personas tienen esta forma de pensar de forma innata, pero también es posible desarrollarla …

El espíritu de un emprendedor

Para tener éxito en el aprovechamiento del big data empresarial, es necesario comprender los problemas que deben resolverse y las nuevas posibilidades que pueden ofrecer los datos. Es por eso que el Data Scientist debe comprender el mundo empresarial en general y la industria a la que está afiliado en particular. El sentido de la comunicación Integrado dentro de la empresa, el científico de datos debe poder comunicar sus descubrimientos técnicos a otros empleados, departamentos de marketing o ventas, por ejemplo. Su función es ayudar a los responsables de la toma de decisiones a tomar las decisiones correctas, proporcionándoles la información necesaria. También debe comprender los problemas de otros equipos y ayudarlos a superar estos desafíos a través del análisis de datos. Para ello, también es importante dominar las herramientas de visualización de datos como ggplot o d3.js.

En conclusión, las habilidades necesarias para un científico de datos son numerosas y específicas. Antes de decidirte por emprender una formación o una carrera en este campo, es necesario determinar si tienes o no el perfil de científico de datos.

Consultar también: Por qué es tan importante el análisis Big Data ; ¿Cómo Big Data está revolucionando el marketing? ¿Se lo ha preguntado?