SDS PODCAST EPISODE 1 WITH RUBEN KOGEL

Cirilo: Este es el episodio número uno con el ex ingeniero químico y ahora mago de la ciencia de datos, Ruben Kogel.

Bienvenido al podcast de SuperDataScience. Mi nombre es Kirill Eremenko, entrenador de ciencia de datos y emprendedor de estilo de vida. Y cada semana, traemos personas e ideas inspiradoras para ayudarlo a construir su exitosa carrera en ciencia de datos. Gracias por estar aquí hoy y ahora hagamos simple lo complejo.

(suena música de fondo)

Hola y bienvenidos al primer episodio del podcast SuperDataScience. No puedo explicar lo emocionada que estoy de finalmente hacer que el espectáculo despegue. He tenido esta idea literalmente durante meses y finalmente hoy estamos comenzando. ¿Y de qué va a tratar todo el espectáculo?

El programa tratará de invitar a los científicos de datos más aspirantes del mundo y hablarles sobre lo que hacen, cuál es su experiencia, qué han aprendido en el pasado en el viaje de la ciencia de datos y qué pueden compartir, qué ideas, qué herramientas y metodologías pueden compartir con nosotros y qué podemos aprender juntos de ellos. Entonces, estoy muy emocionado de que estés aquí desde el principio que estás escuchando el primer episodio. Muchas gracias por ser parte de este viaje. Estoy seguro de que juntos vamos a aprender mucho.

Y estoy muy contento de que este primer episodio haya comenzado con una nota muy alta. Hablé con Ruben Kogel, científico de datos en Udemy. Entonces, si no está familiarizado con Udemy, es la plataforma educativa en línea más grande del mundo.
Actualmente, hay más de once millones de estudiantes que aprenden a través de Udemy, así que si no los ha revisado, definitivamente hágalo. Nuestros cursos son básicamente cualquier cosa que usted

podría imaginar y personalmente también soy instructor en Udemy. Tengo más de veinte cursos allí y cerca de cincuenta mil alumnos. Por lo tanto, es una gran plataforma de aprendizaje y Ruben Kogel es uno de los principales científicos de datos en una de las divisiones de Udemy. Entonces, una división que trabaja en contenido y marketing de contenido y Ruben compartieron algunas ideas muy poderosas sobre lo que hace a diario en Udemy y, además, cómo transfirió su experiencia en ingeniería química a un conjunto de habilidades de ciencia de datos. Cómo tomó ABA y seleccionó específicamente las materias de manera esencial para que pudiera aprender más sobre la ciencia de datos y entrar en ese campo para que pasara de la ingeniería química a la ciencia de datos a través de su MBA.

También habló sobre la comunicación de conocimientos y lo importante que es en el rol de la ciencia de datos. También discutimos muchos otros temas, como la identificación de problemas cuando eres un científico de datos. Qué importante es eso. Cómo combinar ciencia de datos y estrategia de producto. Eso es algo que Ruben hace a diario y aprenderás más sobre eso. Y esa es una habilidad muy poderosa, especialmente si estás trabajando en el área de puesta en marcha. El espacio de puesta en marcha en la empresa se encuentra predominantemente en Silicon Valley o en cualquier otro tipo de ubicación donde haya adquirido esta cultura de puesta en marcha. Luego también hablamos sobre la gestión de un equipo de científicos de datos, por lo que Rubén tuvo una gran experiencia en torno a la cual tenía algunos consejos. Si eres un gerente en ciencia de datos o en ese espacio de análisis, puedes obtener algunos buenos consejos de allí.

Y también, hablamos sobre administrar el flujo de entrada de solicitudes y eso es valioso para cualquier científico de datos sobre cómo administrar el flujo de entrada o las solicitudes y luego Ruben nos dio un ejemplo de su enfoque del tablero Trello, hablamos sobre los mentores de

ciencia de datos y, por supuesto, hablamos sobre muchas herramientas de análisis diferentes. Hablamos sobre (inaudible) sql y complementos que no conozco personalmente. Hablamos de wagon, hablamos de r versus python. Entonces, la pregunta (inaudible) cuál es mejor, cuál es más preferible y hablamos de muchas otras cosas en este podcast, así que estoy seguro de que lo disfrutarán. E incluso mencionamos el análisis de autoservicio, por lo que es un espacio en crecimiento del campo de la ciencia de datos. Entonces, sin más preámbulos, les traigo a Ruben Kogel de Udemy.com y disfruten.

(música de fondo)

Cirilo: Hola chicos, bienvenidos al Podcast. Tengo a Ruben Kogel aquí de Udemy. Súper emocionado por este primer episodio.

Rubén, bienvenido! ¿Como vas?

Rubén: ¡Gracias! Gracias por invitarme. Lo estoy haciendo genial.

Cirilo: Impresionante. Es genial escucharte y para aquellos de ustedes que no saben, conocí a Ruben por primera vez cuando estaba en San Francisco hace un par de meses para la primera conferencia en vivo de Udemy. Y fue bastante emocionante e hizo algunas presentaciones geniales, pero solo para que todos se pongan al día, Rubén, cuéntanos un poco a qué te dedicas. ¿Cuál es su cargo en la empresa y qué hace exactamente por su cuenta?

Rubén: Si seguro. Entonces, soy el gerente sénior de análisis y estrategia en Udemy y básicamente lo que hago es ayudar al equipo de contenido, que es el equipo que analiza los cursos y el contenido en Udemy para averiguar qué hay en el catálogo. , qué hay en la selección de cursos, cuál es la calidad de los cursos, cómo podemos mejorar nuestro catálogo para

hacer más felices a nuestros alumnos. En la práctica, mucho de lo que podría funcionar tiene que ver con: si estamos midiendo lo correcto, si estamos midiendo la satisfacción, si hay datos disponibles para que las personas que están a cargo de traer cursos sepan qué cursos son buenos y cómo hacerlo. medimos también la selección de cursos para que poco a poco podamos construir un catálogo cada vez mejor para nuestros estudiantes.

Cirilo: ¡Impresionante! Eso es genial, así que aplicas técnicas de ciencia de datos para medir ese tipo de métricas. ¿Es eso correcto?

Rubén: Totalmente. Me refiero a que la ciencia de datos viene en diferentes partes de mi trabajo. Así como la parte más básica que es la instrumentación. Así que me empiezan a gustar los esfuerzos de datos en los que quieres asegurarte de que estás midiendo las cosas correctas y esos datos están disponibles. Entonces, para mí significa que estamos midiendo correctamente la satisfacción de los estudiantes y que esos datos están disponibles para que las personas de la empresa puedan acceder a ellos y tomar decisiones al respecto. Hay otro nivel en el que uso la ciencia de datos, luego surgen preguntas más amplias y eso es lo que llamo análisis ad-hoc, por lo que alguien podría preguntarme qué sucede si eliminamos algunos de los cursos de baja calidad de la plataforma. ¿Podemos predecir todo el impacto en la satisfacción de los ingresos? Esa es una especie de pregunta en la que necesita conocer la estructura del problema, pero también proponer algunas predicciones y usar algunas técnicas para evaluar cuál sería el impacto, quizás alguna conferencia y roles, y hay otro tipo de pregunta que puede ser, bueno, sabemos que hay una variedad de cursos y todos tienen diferentes puntajes de calidad y queremos saber qué pasa con estos cursos que impulsan estos puntajes de calidad y sabes que el audio

calidad, la calidad del video y la entrega instruida, por lo que en ese caso, la ciencia de datos entra en juego tanto en términos de estructurar el problema como en la ejecución de algún tipo de análisis estadístico para extraer la importancia de las diferentes variables y volver con una respuesta que dice: "bueno, creo que esta variable es la más importante y sabes que si mueves esa variable en un 1%, tendrá un impacto en todas las calificaciones de los estudiantes, pero en esa cantidad". Así que ese es otro ejemplo.

Cirilo: Eso es muy bonito. Por lo tanto, una especie de dos tipos principales: uno son las métricas de las métricas existentes y cómo puede modificarlas para mejorar la experiencia de los estudiantes, pero otro es donde la primera imagen donde está haciendo una especie de comportamiento análisis y análisis de comportamiento predictivo para ver cómo puede cambiar las cosas para que la experiencia futura sea mejor. Desde mi punto de vista, eso está muy bien y es genial que puedas hacer ambas partes en tu papel como científico de datos en Udemy.

Rubén: Sí, es muy muy genial. En realidad, lo que realmente me gusta de mi rol aquí es que es realmente la interfaz entre la ciencia de datos y una especie de producto de estrategia porque puedo trabajar en un conjunto de datos y ellos llegan a una gran cantidad de análisis de datos, pero al final del día la gente que hable con el vicepresidente de contenido hacia el director de adquisición de cursos y las personas que eran como usted saben que son como en el ancho. Están haciendo el negocio y puedo darles recomendaciones, puedo influir, conocer sus decisiones o influir en el producto con datos, por lo que es esta interfaz realmente genial entre datos y negocios.

Cirilo: Definitivamente y eso es lo que también descubrí que los roles y carreras más interesantes y de mayor impacto suceden al borde de dos campos, ya sea, por ejemplo, podría tomar como Física y química o biología y química, pero

Esos son ejemplos muy exagerados, pero incluso en la ciencia de datos, una cosa es solo hacer análisis y todo el asunto es hacer análisis y al mismo tiempo transmitir los hallazgos y trabajar con aquellas personas que usan análisis. Y solo en eso también, ¿cómo le parece transmitir análisis tan complicados a sus partes interesadas que, como dijo el vicepresidente de contenido? ¿Hay algún enfoque específico que utilice o algún consejo que pueda dar a nuestros oyentes sobre la comunicación de estos hallazgos a las principales partes interesadas de la empresa?

Rubén: Sabes, comenzaré diciendo que cualquier análisis que hagas es inútil e irrelevante si no puedes comunicar los hallazgos a las personas. Por lo tanto, la comunicación es una parte muy importante de ser un científico de datos exitoso. Y, pero hay como elegir... o como huevas que trato de usar cuando comunico cosas, una es traducir conceptos técnicos en algo que la gente pueda entender para que no te guste hablar de conferencias (inaudible). Puede hablar sobre su confianza en los datos o puede seleccionar. Bueno, creo que ya sabes, la predicción o los ingresos predictivos estarán entre esos dos saldos. No tienes que decir que eres el noventa y cinco por ciento

confiado porque eso no agrega mucho valor, pero al mismo tiempo también quiere transmitir precisión en su comunicación y asegurarse de que no arroje respuestas como "sí, creo que deberíamos hacer A porque... Yo creo que es importante que transmita algo como "bueno, miro los datos y si hacemos A, podemos aumentar los ingresos en esta cantidad y tal vez solo un poco de un individuo en términos de lo que creo que podemos aumentar los ingresos, pero esto es lo que está haciendo el análisis. Entonces, es como este equilibrio entre ser preciso y mostrar que has hecho tu tarea y mostrar que esto es un dato escrito pero al mismo tiempo traducir

a las palabras del profano y eliminando cualquier tecnicismo que no agregue muchos valores a su mensaje.

Cirilo: Totalmente de acuerdo. Tengo un ejemplo clásico sobre eso, con qué frecuencia las partes interesadas, especialmente las partes interesadas principales, son muy escépticas sobre los tamaños de muestra, por ejemplo, así que si realiza un análisis en un tamaño de muestra de 157 o 300, podrían ser propensos a decir que en realidad queremos una muestra. tamaño de 10,000 pero usted, como científico de datos, sabe que el tamaño de muestra que ejecutó es significativo. Es estadísticamente significativo, por lo que es importante transmitir estos hallazgos de manera que cuando tenga confianza en sí mismo, no tenga que entrar en todos esos detalles para explicar exactamente la metodología detrás de esto, pero en realidad solo transmita la confianza en el forma en que presentas, la forma en que posicionas tu análisis tan totalmente de acuerdo en eso una.

Lo interesante que me gustaría preguntarte y probablemente muchos de nuestros oyentes tengan curiosidad es: ¿puedes contarnos un poco de tu experiencia? Entonces, ¿cómo llegaste a ser científico de datos y progresaste? Por supuesto, progresas más en tu carrera y ahora eres el jefe de análisis en ese departamento, pero originalmente, ¿cómo entraste en la ciencia de datos porque este es un campo bastante nuevo y en el pasado no se enseñaba en la universidad, entonces, ¿cómo llegaste aquí y cuáles son los pasos?

Rubén: Es una pregunta muy interesante porque tuve un camino muy serpenteante. No comencé en absoluto en datos. Mi experiencia fue en física aplicada y luego cambié a ciencia de materiales y fui ingeniero químico durante muchos años.
Estaba tratando con datos, pero no con el tipo de análisis que usted hacer cuando eres científico de datos. Era mucho más rudo y mucho más sofisticado y lo que sucedió es que en realidad estaba

buscando como la transición a algo diferente. Fui a la escuela de negocios y normalmente no piensas en una escuela de negocios como un lugar donde aprendes ciencia de datos, pero tuve esta oportunidad de usar cursos de estadística y uno como el curso de minería de datos que de repente me enamoré del campo. y cuanto más aprendo al respecto, más me intrigaba y realmente aprendí la teoría en la escuela de negocios y tuve esta oportunidad de venir a trabajar a

Udemy y apliqué algunos de mis conocimientos y así fue como comencé mi carrera. Así que es una buena razón y también fue un contraste bastante marcado con lo que estaba haciendo antes.

Cirilo: Definitivamente. Es un gran salto y, como dices, uno no esperaría que aprendieras todas las habilidades de datos necesarias en escuela de negocios, pero supongo que eligió las materias correctas y eso es un gran testimonio de lo lucrativo que es el campo de la ciencia de datos y cuéntenos las habilidades que desarrolló como ingeniero químico o trabajando en ese campo. ¿Hay alguna forma de aprovecharlos actualmente porque la ciencia de datos proviene de todas las áreas diferentes? Algunas personas provienen de clases de actuación, algunas personas pueden provenir de economía o finanzas, pero con antecedentes en química, ¿hay alguna habilidad o mentalidad particular que pueda compartir con nosotros que pueda aprovechar con su trabajo actual como científico de datos?

Rubén: Sí. Totalmente. Sorprendentemente, no son las habilidades de datos las que aprovecho de mi experiencia en ingeniería química. Es más la resolución de problemas en las habilidades de comunicación. Creo que cualquier trabajo de ingeniería tiene un componente importante de resolución de problemas y solución de problemas, por lo que estaba haciendo mucho de eso en mi trabajo y realmente me obligó a pensar en cosas como un enfoque muy sistemático para desglosar un problema. Proponer una hipótesis y creo que esas hipótesis

y ser extremadamente sistemático y organizado y estructurado acerca de mi pensamiento. Así que definitivamente aprendí eso de mi experiencia en ingeniería. Lo único que aprendí es, como mencioné, habilidades de comunicación. Como ingeniero, especialmente en mi puesto, estaba haciendo muchas cosas como la gestión de cuentas. También tuve que traducir una gran cantidad de experimentos y resultados técnicos muy complejos en algo que la gente de negocios pudiera entender y la capacidad de resumir conceptos y nociones complejos en unos pocos, ya sea una diapositiva o un correo electrónico que conozca y un borrador preparado. que realmente resume las ideas, es muy importante que algo que aprendí en mi trabajo de ingeniería y algo que me emocionó también en mi actual

carrera profesional.

Cirilo: Suena genial y definitivamente las habilidades de resolución de problemas son algo muy valioso cuando se trata de desafíos de ciencia de datos, pero por lo que dice, deduje que sus habilidades de comunicación habían jugado un papel muy importante en términos de su éxito y ¿Cómo lo recomendaría? Porque, en mi opinión, muchas veces, cuando las personas se inician en el campo de la ciencia de datos ahora que se está volviendo cada vez más popular, a veces pueden quedarse en suspenso en ciertos roles en los que desempeñan algunas funciones. análisis, pero no tienen la exposición para ir y compartir sus conocimientos con las partes interesadas. Solo estaban realizando ciertas consultas SQL o ciertos procedimientos analíticos, pero en realidad no tienen la oportunidad de comunicar ideas. Entonces, ¿cómo sería su sugerencia desde la parte superior de su cabeza para las personas o nuestros oyentes que podrían estar en esa situación para que de alguna manera comiencen a desarrollar esas habilidades de comunicación?

Rubén: Sí. Creo que hay dos formas de hacerlo: una es, incluso en su interacción diaria con los clientes. Pensaría en el análisis de las ciencias de datos, sabes que tienes clientes dentro de la empresa, ya sean clientes técnicos o clientes comerciales, las personas te hacen preguntas o te piden que realices un análisis de datos para darles una respuesta, así que cada vez que interactúas con tus clientes , siempre puede hacer un buen esfuerzo adicional y estructurar su respuesta no solo como el resultado de un análisis de regresión o como una consulta SQL, sino que podría tratar de explicar por qué cree que esto es lo correcto, qué significa eso en la práctica , ¿cuál sería su recomendación para que siempre impulse la ruta de los resultados, no solo el resultado técnico, sino que lo empaquete de una manera que muestre que pensó en las implicaciones y el significado del análisis, así que eso es una cosa?

La segunda cosa es que también siempre recomiendo a las personas y eso es cierto para las personas que trabajaron para mí o las personas que podrían trabajar en otro puesto que siempre que se les presente un problema, siempre es una buena práctica tratar de profundizar exactamente en el problema. persona está tratando de resolver. Porque a menudo, alguien dejará comentarios como "oye, ¿puedes extraer estos datos o puedes crear un tablero para eso o puedes hacer un análisis sobre esto?" como si realmente estuvieran tratando de resolver el problema y es posible que no te digan cuál es. el problema tratando de resolverlo, así que si interactúa con esa persona o cliente, intente realmente llegar al fondo del problema que están tratando de resolver. De repente, estás comenzando una conversación sobre lo que estás tratando de resolver, qué puedes aportar además del análisis aprobado, cómo puedes ayudarlos a enmarcar tu problema y, de repente, estás involucrado en la comunicación, estás comprometido en como

expresar un problema, dividir y elegir algunos problemas de datos más elementales y volver con un
solución que abordó y realmente en el problema de la línea. Así que esa es una forma en la que puede empujar los límites de su trabajo actual y realmente expandirse para ofrecer más valor que se basa en esta comunicación y se basa en comprender el problema subyacente. Así que ese es el número uno.

Creo que el número dos también es, en términos generales, los analistas tienen esta capacidad única de analizar los datos, obtener algunas ideas que nadie más en la organización tiene y, por lo tanto, como analista, también tiene la oportunidad de comenzar a abordar los problemas. que otras personas también pueden tener o algunas personas pueden no haber pensado y puede crear valor siendo un poco proactivo sobre lo que cree que deberíamos investigar, lo que cree que sería un análisis útil con algunas ideas útiles para otros equipos. Entonces, también existe esa oportunidad porque nadie más puede realmente ver los datos, usted es la única persona que tiene acceso a los datos y también puede sacar a la luz los conocimientos.

Cirilo: Hermoso. Me encanta. Especialmente el: cuál es exactamente su problema, que es una de las mejores habilidades para tener como científico de datos para ayudar a las personas a identificar el problema porque a menudo solo obtienen datos en lugar de identificar realmente el problema e independientemente del nivel en el que se encuentre como dato. científico o analista de datos, finalmente, esa habilidad es la que lo impulsará y combinándola con lo que dijo sobre el enfoque proactivo, ahí es donde se convierte en el médico de la organización y camina y mide lo que está saliendo mal y cómo puede ayudar arreglalo. Así que definitivamente estoy de acuerdo con esos dos. Esos fueron algunos de los

habilidades muy poderosas para tener. Y mencionaste que estás a cargo de algunas personas. ¿Puede contarnos más acerca de cuántas personas tiene a su cargo y cómo llegó a administrar el equipo de analistas o científicos de datos? ¿Cuáles son los retos a los que te enfrentas en el día a día?

Rubén: En este momento, mi equipo se reduce a una sola persona, por lo que en el HayDay teníamos un equipo más grande, pero las cosas se movieron bastante rápido. En Silicon Valley, tengo una persona informando y contratando al menos a una persona por el momento. Bueno, lo que encuentro que es más desafiante en términos de gestión y crecimiento de analistas o científicos de datos es que a menudo existe esta tensión entre tratar de complacer a sus clientes internos y tratar de hacer felices a tantas personas como sea posible en el menor tiempo posible. . Y tratar de generar valor a largo plazo y trabajar en proyectos similares a más largo plazo.
La forma en que pienso en un científico de datos es como si fueras el final de la cadena. Hay muchas personas en la organización que sabes que administran proyectos, les piden a otras personas que hagan cosas y sabes que eventualmente, cuando reúnes todas las contribuciones, obtienes un producto, pero cuando estás en ciencia de datos, no lo haces. pedirle a otras personas que hagan cosas o es muy raro, como normalmente, eres la última persona a la que preguntan y, por lo tanto, muchas personas acuden a ti y terminan teniendo un montón de solicitudes y gestionando el flujo de solicitudes y gestionando las diferentes cosas en las que están trabajando. ser muy desafiante a menudo como un nuevo analista, tienden a atraer sus palabras, tal vez como los proyectos a más corto plazo, piensan que la última persona que se envía a hacer es como "oh, sí, lo haré de inmediato" y al final detrimento de trabajar en un proyecto más grande a largo plazo y más impactante, así que ese es uno de los desafíos.

Cirilo: Eso es definitivamente cierto y eso fluye hacia el arte de decir no a las personas que acudirán a usted como científico de datos, especialmente cuando está a cargo de un departamento y cuando tiene algunos éxitos encontrará aún más de los otros departamentos. de la empresa y de las personas y partes interesadas que acuden a usted con solicitudes. Entonces, ¿cómo dices que no? ¿Cómo le dices a la gente que, “oye, tu proyecto es genial y me encantaría trabajar en él pero al mismo tiempo tengo otros compromisos? ¿Cuáles son tus consejos al respecto?

Rubén: Sabes que en realidad no dices que no. La verdad es (inaudible) que dejas muy claro cuáles son tus prioridades y una de las herramientas que usamos aquí en Udemy es un tablero de Trello. Así que Trello es una herramienta de actividad que es como post-its virtuales. Esencialmente, te permite mostrar un tablero público de lo que estás trabajando, cuáles son las etapas de los diferentes proyectos, cuáles son las partes interesadas de los diferentes proyectos y si alguien te pide que hagas un análisis, puedes decir que está bien, no hay problema, ¿puedes? simplemente registras en mi pizarra y rápidamente se dan cuenta de que su tarjeta es una de otras 50 tarjetas y ahora hay un proceso de priorización para que el equipo comience a trabajar en ella. Así que ese es uno en el que estamos trabajando.

La segunda parte es como, realmente lo que quieres hacer es como si no quisieras trabajar en cada una de las solicitudes de joyas y venderlas. Lo que quieres construir es una infraestructura escalable. Solo crea análisis escalables. Lo que eso significa es que en lugar de responder la misma pregunta una y otra vez, o en lugar de obtener datos para todos en la empresa, crea herramientas de autoservicio. Usted crea tablas, tableros, interfaces de usuario web que permiten a las personas acceder a los datos que necesitan para que no le pidan que haga los datos y extraiga el análisis de datos.

más. De modo que puede liberar algo de su ancho de banda para trabajar en el proyecto más interesante.

Cirilo: Impresionante. Me encanta. El análisis de autoservicio se ha convertido en un concepto cada vez más popular en el mundo de la ciencia de datos. Específicamente por esa razón: para liberar a los científicos de datos y empoderar al usuario final para que haga lo suyo y, por lo tanto, ¿cuáles son las herramientas que usa en Udemy, si puede divulgar esto para el lado analítico de autoservicio?

Rubén: La base del análisis de autoservicio es la creación de un conjunto de tablas de resumen que tienen toda la información relevante que cubre el noventa por ciento del caso de uso. Lo que quiero decir con eso es que, por ejemplo, las preguntas (inaudibles) que vienen en Udemy son como "Oh, ¿tienes una lista de cursos que se han publicado este mes o cuál es el ingreso total de los cursos que se publicaron?"

publicado el mes pasado o puede buscar un instructor en particular y ver cuántas inscripciones o revisiones tiene en sus cursos. Entonces, hay un conjunto limitado de preguntas que aparecen una y otra vez y, en lugar de crear consultas o cada vez que extrae la información, simplemente crea una o dos tablas que tienen toda esa información resumida para que las personas puedan acceder a los datos directamente. adquiriendo la tabla por secuela o buscándola en el tablero. Entonces, en la práctica, lo que haríamos es tener todo nuestro flujo de información de datos en su (inaudible) por Amazon y eso le permite construir tablas encima de tablas sin procesar, por lo que tenemos lo que llamamos tablas de resumen construidas sobre las tablas sin procesar. O pueden alimentar tableros entre la interfaz Ui de chartio usada y pueden usar el chartio para extraer cualquier información que necesiten.

Cirilo: De acuerdo, corrimiento al rojo y chartio. Sí. Esas son las dos herramientas. Una muy buena respuesta. Almacena Amazon en un WS para el almacenamiento de datos, ¿correcto?

Rubén: ¡Uhhh!

Cirilo: ¿Y cómo encontraste eso? ¿Ha sido ese un motivo de transición de su organización o siempre ha sido así?

Rubén: No. Ha sido una transición. Hace aproximadamente un año y medio solíamos tener solo nuestro propio servidor MySQL tradicional y luego comenzamos a explorar el desplazamiento hacia el rojo. (inaudible) corrimiento al rojo y vimos que era mucho más poderoso en términos de hacer análisis porque, como la base de datos típica de My SQL, está optimizada para escribir, pero honestamente no para crear nuevas tablas. El corrimiento al rojo está optimizado para hacer muchas uniones, muchos análisis y lectura de datos. Por lo tanto, hemos estado usando el desplazamiento hacia el rojo durante el último año y medio y hemos escalado el tamaño de nuestro clúster a medida que crecen nuestros datos y nuestro equipo de análisis y nos ha estado sirviendo bastante bien en realidad.

Cirilo: Escuché muchos comentarios sobre que en AWS puede escalar según sus necesidades y esa es una de las mayores ventajas de AWS que, a medida que su organización crece y necesita más capacidad para empoderarse y escalar, no tiene que comprar. eso por adelantado y también hay mes a mes u otro tipo de planes ahí. Muy conveniente. Por lo general, la única organización de respaldo que también tiene esto en común es una organización que se ocupa de los datos de los clientes, los datos de cara al cliente, como Udemy. Tendrían ciertos problemas regulatorios, quizás ciertos problemas regulatorios con la subcontratación del almacenamiento de los datos en la nube o en sistemas externos como AWS. ¿Enfrentó algo de eso cuando estaba haciendo la transición?

Rubén: No realmente en el sentido de que los datos siguen siendo seguros y en realidad, en los estados unidos, los únicos datos que están fuertemente regulados son los datos de salud, por lo que si tiene el tipo de regulaciones de salud que dificultan el trabajo con la nube o tiene que trabajar con proveedores certificados que realmente pueden manejar el principio -ups la regulación sobre cómo maneja los datos de salud para el resto AWS sirve a una gran variedad de empresas emergentes y todas tienen datos confidenciales, pero usted sabe que están configurados para manejar los datos confidenciales correctamente, por lo que no hay mucha preocupación al respecto. Lo único es que a veces desea que los datos se utilicen para aplicaciones web, en cuyo caso sabe que leer desde el desplazamiento hacia el rojo no es la mejor manera de completar el campo en su sitio web o aplicación web, por lo que es posible que necesite un segundo representante de datos que sea más rápido. para leer según el almacenamiento del calendario.

Cirilo: Entonces, comenzamos a profundizar en la herramienta para poder pasar ahora a las herramientas de administración como Trello, lo cual me pareció bastante interesante: cómo lograr que alguien publique un post-it en Trello en su tablero y luego se dieron cuenta de que "oye, hay proyectos". eso se va a priorizar y, en algún momento, es posible que el suyo no se haga muy rápido" y ahora pasamos al desplazamiento hacia el rojo. ¿Cuáles son las otras herramientas que utiliza a diario en su función de análisis?

Rubén: Las dos tecnologías que utilizo todo el tiempo son SQL o, en ese caso, SQL, que es la base para el corrimiento al rojo y la R, por lo que hago gran parte de mi análisis en R. Básicamente, ese es el lenguaje que aprendí y estoy muy familiarizado con él. Algunas otras personas en Udemy usan Python. Realmente varía. Pero creo que una de las herramientas es que son muy convenientes y flexibles y esas son las herramientas de los científicos de datos hoy en día.
Ya sea python o r tienen como los paquetes e instaladores

fuente abierta. Crece como una comunidad muy activa, por lo que son típicos para el análisis de datos. En términos de la herramienta que uso donde realmente hago mi SQL, hay una muy buena compañía llamada Wagon. Esto tiene un mejor producto, pero en este momento tienen el mejor editor de SQL, especialmente para SQL posterior (inaudible). Es realmente genial porque puedes organizar tus consultas en diferentes carpetas, todo siempre se guarda en la nube para que no se pierda nada y es muy parecido a una interfaz ordenada y es muy fácil de usar.

Cirilo: Impresionante, ¿así que eso fue Wagon?

Rubén: Vagón, sí.

Cirilo: ¿Cómo se deletrea eso?

Rubén: VAGÓN

Cirilo: Hermoso. Así que no he oído hablar de eso antes. Algo para definitivamente revisar. Y si. Muy interesante cómo su organización tiene una división entre r y python. Supongo que en el mundo de las empresas emergentes es más común, pero en las organizaciones más grandes que han existido durante un tiempo y que tienen mucho legado detrás de ellas, por lo general, los analistas no pueden darse el lujo de poder elegir entre los dos. Definitivamente, R es algo en lo que muchos de nuestros oyentes están interesados porque sabes que tal vez compartas un par de paquetes o un par de técnicas que usas más comúnmente cuando codificas en R.

Rubén: Sí, totalmente, así que probablemente te decepcionaré a ti y a tus oyentes. No uso muchos paquetes diferentes y técnicas avanzadas. Tiendo a apegarme a la R básica, de hecho, ni siquiera corro como lo hacen los demás. Solo uso la R básica, no uso la meseta GG. Solo uso como los gráficos básicos de R. pienso

parte de la razón es que no veo a R como una forma de producir informes o análisis extremadamente sofisticados. Uso R principalmente para extraer la información que necesito y ejecutar como lo básico, el análisis básico, por lo que lo típico que haría en R es importar un archivo csv, leerlo, hacer un literal limpieza, aunque por lo general prefiero hacer mi limpieza en SQL porque no creo que debas hacer la limpieza en absoluto.
Creo que realmente la limpieza debería ocurrir aguas arriba y solo inicias R para hacer una exploración de datos, ya sabes, crear algunos gráficos y hacer algunos análisis estadísticos. Por lo general, ejecutaría el análisis de regresión que uso para eliminar para hacer muchas regresiones, uso (inaudible) para usar un modelo aleatorio o entender la importancia relativa de las diferentes características y para su modelo, así que es como mi uso de R.

Cirilo: Maravilloso. Estoy totalmente de acuerdo con eso. R es definitivamente una herramienta muy poderosa y cada analista, cada científico de datos tiene derecho a usarla de la manera que prefiera y usted usa r en un enfoque muy esbelto que también puede ser muy poderoso. La pregunta: la pregunta del millón de dólares sería R versus Python. ¿Cuáles son tus pensamientos y por qué terminaste eligiendo R?

Rubén: Mi respuesta podría decepcionarte de nuevo. Como si nunca hubiera elegido realmente entre los dos. Empecé con R y se ajusta a todas mis necesidades. Hablé con un par de personas que usan python y rápidamente me di cuenta de que el tipo de análisis que estaba haciendo, ya sabes, el análisis fuera de línea, la creación de un modelo fuera de línea, tratar de comprender los impulsores de una métrica en particular, Python sería más complejo y no agrega mucho valor. Lo que quiero decir con eso es que en mi trabajo diario, no construyo productos de datos reales. No trato de implementar un modelo predictivo en nuestra infraestructura y, por lo tanto, no necesito estar en

Pitón. Todo lo que hago es descargar algunos datos generalmente del desplazamiento hacia el rojo y trato de construir algún modelo para comprender qué está impulsando esta métrica hacia arriba o hacia abajo y para este tipo de caso de uso, por lo que leí, R es como el más simple y más directo. ca de anunciar los datos y también es como el idioma que conozco.

Cirilo: Hermoso. Definitivamente, ese también es el caso en muchas situaciones cuando comienzas con uno y luego continúas con ese si se adapta a tus necesidades y cuál es el punto de cambiar.
Totalmente aprecio eso. Bien. Eso es realmente genial y entramos en detalles sobre las herramientas que usas y algunas técnicas. Me encanta esa parte de la conversación. Pasemos a un poco más, algunas de las cosas más suaves. Por ejemplo, puedo ver que ha cambiado y ha hecho esta transición a la ciencia de datos desde la química y nunca ha mirado atrás. Fue como superarlo y progresar en su carrera, hacer crecer a otros científicos de datos en su equipo y actuar como mentor. Pero en el camino, ¿tuviste alguna influencia que te ayudó a convertirte y a perseverar en esta carrera de ciencia de datos y tal vez algunos mentores que podrías haber tenido o pasatiempos o algunos eventos que cambiaron tu vida o incluso algunos artículos o algo que realmente te influenció y ayudó a lo largo de esta trayectoria profesional como científico de datos.

Rubén: Tuviste una gran curva de aprendizaje. Nunca he ejercido como científico de datos y, de hecho, no he tenido un mentor en Udemy, así que tuve que resolver muchas cosas por mí mismo o preguntarle a la gente de afuera. y así disfrutar. Realmente animo a la gente a buscar mentores.
En mi caso, tenía un buen amigo mío que tenía un poco de ventaja sobre mí. Comenzó a hacer ciencia de datos durante cinco años antes que yo. Es alguien que tiene opiniones firmes, pero también es muy reflexivo sobre el

diferentes enfoques y elecciones que estaba haciendo y, a menudo, ya sabes, tuve conversaciones con él. Tomamos como un café regular donde intercambiamos sobre tecnologías y técnicas. Encontré esta interacción con él muy útil. Fui a diferentes conferencias, aunque diría que todas fueron útiles, pero la que realmente me gustó fue la de Airbnb en San Francisco, la conferencia llamada al aire libre que realmente aprecio y también seguí algunos blogs y boletines. Hay un boletín que me gusta mucho llamado datascienceweekly.org y es una colección de artículos interesantes sobre 10-12 artículos interesantes cada semana. No los leo todos, solo elijo el que parece interesante y si sabes si paso el primer párrafo y es realmente interesante, entonces es para leer y luego, lentamente, construyo un catálogo de consejos. Pensamientos suaves que creo que son muy útiles y, en particular, no recuerdo si es a través de este boletín o tal vez algo como Linkedin en línea. Encontré a alguien publicando en un artículo antiguo de Leo Breiman. El tipo que creó los bosques aleatorios que publicó en 2001 y que realmente resonó conmigo, por lo que realmente alentaría a cualquiera que esté considerando la ciencia de datos o comenzando en ciencia de datos o alguien que esté incluso más avanzado en su carrera, lea este artículo porque para mí realmente ha expresado exactamente lo que siento sobre la tensión entre las estadísticas y el aprendizaje automático y la tensión que siento entre la construcción experimental

modelos y modelos predictivos y lo hace de una manera muy convincente y ordenada. Entonces, el artículo se llama Modelado estadístico: las dos culturas por Leo Breiman, que se publicó en ciencia estadística en 2001. Me parece que es el mejor para leer para obtener realmente algunos fundamentos en ciencia de datos.

Cirilo: Guau. Maravilloso. Definitivamente no he oído hablar de ese artículo, pero definitivamente lo revisaré. Suena como una gran lectura y eso es de Leo Breiman. Definitivamente lo incluiré en las notas del programa. La siguiente pregunta que tengo es si pudieras compartir con nosotros alguna victoria reciente en ciencia de datos, victorias que hayas tenido en tu departamento en Udemy.

Rubén: Dos de ellos vienen a la mía. Mi equipo ha sido responsable de crear el nuevo filtro de spam en Udemy, por lo que es el filtro que determina si una revisión es confiable o no y el antiguo filtro de spam se creó en un conjunto de reglas. Así que ni siquiera era un bayes ingenuo, era solo un conjunto de reglas que tenían mucho sentido en ese momento, pero con el tiempo sabes que la gente aprende a eludir las reglas. Aprendieron que el filtro lógico y que había mucho contenido de spam no confiable en Udemy y mi equipo abordó este problema y pudimos mejorar la precisión del filtro de spam en un factor 8X, por lo que es una gran victoria para el equipo, pero principalmente para la empresa en su conjunto.

Ese fue uno de ellos. El segundo que me viene a la mente fue un análisis en movimiento y ad-hoc que hicimos hace un par de semanas, por lo que eres un instructor en Udemy. Usted sabe que cambiamos la estrategia de precios y, por lo tanto, hubo un cambio en el comportamiento de los estudiantes y mi equipo analizó el precio final, el precio de lista, el descuento, todos estos influyen en la decisión y si podemos construir un modelo cómo lo haría el estudiante. reaccionan a ellos a diferentes estrategias de precios, por lo que era un modelo muy simple y no compraba nada complejo realmente sofisticado y tenía la capacidad de explicar los datos que hemos observado en el pasado y, por esa razón, era poderoso tanto porque tiene experiencia poder pero también fue construido simple para que la gente pudiera entender lo que significaba.

Cirilo: Es genial ese filtro de spam de precisión 8X. Esa es una mejora significativa y un modelo para el comportamiento de diferentes contenidos que también es muy interesante. Como dices, soy un instructor en Udemy, puedo ver el backend de estas cosas y cómo se ejecutan en el backend, definitivamente puedo ver cómo se producen los cambios y cómo crece la plataforma, por lo que es muy emocionante saber que tú' estás detrás de todos esos cambios. ¡Eso es realmente genial!

¿Qué es lo que más te gusta de ser científico de datos? ¿Qué es lo que te emociona para levantarte e ir a trabajar por la mañana y lo emociona para hacer tu trabajo y qué es lo que te impulsa a seguir adelante?

Rubén: Creo que lo más emocionante es el desafío intelectual. Es el hecho de que siempre te enfrentas a problemas nuevos y sin resolver y eres la persona a la que se le pide que resuelva esos problemas. Y a veces es sólo un problema de datos. A veces es más complicado que eso. A veces tienes que estructurar el problema y plantear las preguntas de datos correctas, resolverlas y darles respuestas. Ese desafío intelectual constante es realmente lo que me motiva.

Cirilo: Y para nuestros oyentes, desde su perspectiva, desde lo que ha visto, lo que ve actualmente en el campo de la ciencia de datos, cómo lo ha visto evolucionar desde que se unió a las filas. ¿Hacia dónde crees que va este campo? ¿Para qué crees que deberían prepararse nuestros oyentes en el futuro? ¿En qué deberían enfocarse? ¿Qué habilidades deberían desarrollar o qué técnicas deberían pensar al respecto o, en general, hacia dónde cree que se dirige todo este campo?

Rubén: Sí. Buena pregunta. Es difícil predecir realmente hacia dónde se dirige esto y, si me gusta una visión generalizada, puedo mencionar un

algunas tendencias y puedo mencionar también algunas áreas donde creo que las personas realmente pueden marcar la diferencia y mostrar gusto y agregar valor. Así que en este momento hay muchas conversaciones sobre la plataforma de datos. Entonces, no es solo que tenía una infraestructura de datos, está extrayendo los datos y está construyendo algunos modelos para trabajar con algunos conocimientos, sino que también existe la idea de que para operar un equipo de datos eficiente y un equipo de análisis eficiente necesita tener una mejor plataforma que permite a los científicos de datos implementar experimentos, ejecutar experimentos, construir y validar modelos rápidamente. Entonces, está ese aspecto de construir la canalización y luego los flujos de trabajo que permiten a las personas escalar básicamente el análisis, por lo que esa es una dirección en la que van las cosas y, obviamente, esto también va en la dirección de un poco de especialización. Solía ser que este ingeniero que podía construir una base de datos porque también extraía datos, ejecutaba algunos análisis estadísticos y mostraba los resultados al jefe de marketing y ahora usa cada vez más reglas siendo un poco más especializado, ahora tienes personas que se especializan en almacenamiento de datos e infraestructura de datos y hay personas que se especializan en plataformas de datos, personas que se especializan en algoritmos, personas que se especializan en análisis es parte de la ciencia de datos. Entonces, creo que es importante comprender todos estos roles y definitivamente hay algunos, definitivamente son algunas de las tendencias en la industria. Paralelamente, creo que lo que tenemos que recomendar es que creo que es muy importante que las personas desarrollen experiencia tecnológica porque eso tiene mucho valor. Si eres capaz de codificar en Python, en R y tal vez incluso sabes agregar un poco de Java y entiendes todas estas tecnologías, es extremadamente poderoso pero al mismo tiempo, advertiría a las personas que no se apeguen demasiado a cier técnicas de aprendizaje automático. Siempre habrá gente

que actúan se especializan en aprendizaje profundo y recurren a su antiguo (inaudible) sabes que a menos que seas una de esas personas, realmente no necesitas ir en esa dirección. Tampoco es necesario que aprenda todos los algoritmos diferentes. Creo que es más importante comprender qué hacen las diferentes técnicas y realmente profundamente, como tener una comprensión profunda de las estadísticas y cómo usas diferentes cosas en diferentes casos y tener la capacidad de aprender y luego aplicar el modelo correcto o la técnica correcta para el problema correcto.
Por lo tanto, es más importante poder trazar la técnica correcta para el problema correcto en lugar de conocer todas las técnicas y algoritmos posibles que existen en este una.

Cirilo: Un consejo muy, muy poderoso allí. Simplemente lo resumo para todos los oyentes y solo para mí también. Está observando una tendencia en la que los científicos de datos se están volviendo más maduros como un tipo de industria, tipo de trabajo y, por lo tanto, algunos roles se están volviendo más especializados. Entonces, supongo que es una buena idea que el analista y el aspirante a científico de datos comiencen a buscar lo que más les interesa y eventualmente terminen haciendo algo que les apasione en este campo y también que desarrollen una profunda experiencia tecnológica en un una amplia gama de herramientas y técnicas es muy importante porque no querrás quedarte estancado usando solo esa técnica porque este campo está en constante evolución y quieres ser capaz de adaptarte y aprender nuevas habilidades sobre la marcha, como dicen. Creo que es un consejo muy poderoso y sé que ya has recomendado un gran artículo por lo que parece del creador de Random Forest. ¿Hay algún libro, un libro que podría recomendar a nuestros oyentes si tuvieran tiempo para aprender algo y mejorar su ciencia de datos?

carreras y habilidades. ¿Cuál podría ser el único libro que crees que deberían leer?

Rubén: Esa es una buena pregunta porque en realidad nunca aprendo ciencia de datos de un libro. Lo aprendí en la escuela, lo aprendí haciendo o mirando sitios web y foros, sin embargo, hay un libro que influye en mi forma de pensar sobre el análisis de datos y realmente cimentó mis ideas sobre la adquisición, la correlación, ¿cómo se puede torturar? datos para ver ciertas cosas, y cómo eso podría estar mal y cómo puede mirar los mismos conjuntos de datos y cómo puede llegar a conclusiones. Y ese libro es en realidad The Signal and the Noise de Nate Silver. Es un libro popular y es un libro que obviamente es muy técnico, pero creo que las ideas en ese libro son extremadamente poderosas y nuevamente es esta idea de que sabes que la ciencia de datos no es solo un conjunto de técnicas y herramientas, también es una forma de pensar. sobre el problema y si no piensa correctamente sobre el problema, no importa que tenga las técnicas con las que terminará, como ideas y conclusiones incorrectas. Es importante tener una comprensión profunda de lo que está tratando de lograr, cómo le gusta abordar un problema, cómo lo ve correctamente y luego usa las técnicas y herramientas adecuadas. Entonces, por esa razón, recomendaría The Signal and the Noise de Nate Silver.

Cirilo: Hermoso. No he leído el libro yo mismo. Definitivamente eso va a mi lista de libros que recogeré en un futuro próximo.
La señal y el ruido de Nate Silver.

Ha sido un placer, Rubén. Para nuestros oyentes, ¿dónde pueden encontrarte? ¿Cómo pueden contactarlo, seguirlo en cualquier red social, cualquier sitio web, dónde pueden ponerse en contacto con usted?

Rubén: Creo que lo más fácil es que se ponga en contacto conmigo en LinkedIn. Estoy trabajando en un blog y un sitio web. Todavía no está disponible, pero el enlace definitivamente estará en mi perfil de LinkedIn. Así que esa es una de las mejores maneras.

Cirilo: Definitivamente y también incluiré el enlace en las notas del programa. Si hay alguna actualización, también la incluiré en las notas del programa. Muchas gracias, Rubén. Realmente aprecio que vengas al programa y compartas tus ideas. Creo que este ha sido un gran día. Maravillosa y esclarecedora conversación. Muchas gracias por venir.

Rubén: Sí. Fue un placer. Gracias por invitarme.

Cirilo: Así que ahí van chicos. Ese fue Rubén Kogel. Espero que hayas disfrutado el espectaculo. Puede obtener las notas del programa en www.superdatascience.com/1. Allí también puedes dejarme a mí o a Rubén un comentario en la sección de comentarios en la parte inferior. Haga una pregunta o díganos lo que piensa. Además, si disfrutó del programa, asegúrese de compartirlo con sus amigos y compañeros de trabajo para que pueda ayudarnos a correr la voz.
sobre el podcast de SuperDataScience y espero verlos la próxima vez. Hasta entonces, feliz análisis.