

Tworzenie zespołu

Nastawienie na naukę danych jest wyzwaniem zarówno organizacyjnym, jak i kulturowym. To nie jest tak proste, jak zatrudnienie kilku analityków danych. Chcesz, aby cała organizacja myślała o Twoich danych w kreatywny i interesujący sposób. Analitycy danych pomogą Ci przeanalizować dane, ale mogą nie być najlepszym źródłem nowych informacji. Jak wspomniano w Części 6, powinieneś myśleć o nauce o danych jako o przedsięwzięciu zespołowym - małe grupy ludzi z różnych środowisk eksperymentują z Twoimi danymi w celu tworzenia wiedzy. To jest metoda naukowa w nauce o danych. To empiryczny proces eksploracji. Zadasz dobre pytania, zbierzesz dowody i spróbujesz wyciągnąć wnioski. Zamiast szukać kilku osób, które mogą zrobić wszystko, pracuj z istniejącymi ludźmi, którzy potrafią zrobić większość rzeczy. Jednym z modeli, który moim zdaniem dobrze się sprawdza, jest podzielenie zespołu zajmującego się analizą danych na trzy role:

- Prowadzenie badań
- Analityka danych
- Menadżer projektu

Widziałem te trzy role w różnych organizacjach. Mogą mieć różne nazwy, ale starają się osiągnąć to samo: zadać dobre pytania, zbierać dowody i próbować uzyskać wgląd. Każdą z tych ról omówimy bardziej szczegółowo w kolejnych sekcjach, a następnie omówimy, jak współpracują one w grupie.

Zadawanie dobrych pytań z kierownikiem ds. badań

Albert Einstein miał powiedzieć, że gdyby miał godzinę na rozwiązanie problemu, pierwsze 55 minut spędziłby na szukaniu właściwego pytania. Pytania są kluczem do odkrycia. Właśnie dlatego są niezbędne dla zespołu zajmującego się badaniem danych. Pytania są najważniejszymi czynnikami napędzającymi Twoje nowe spostrzeżenia. Kluczową częścią „nauki” w nauce o danych jest znalezienie właściwego pytania. Masz już dane. Twoja organizacja prawdopodobnie zbiera więcej danych niż potrzebujesz. Pod wieloma względami już zebrześ odpowiedzi. Teraz musisz zadać właściwe pytanie. Dla większości organizacji nie jest to łatwe zadanie. Nadal pracujemy dla firm, które pragną odpowiedzi, ponieważ odpowiedzi kończą dyskusję. Wychodzisz ze spotkań i zaczynasz wprowadzać coś w życie. Dlatego organizacje mają tendencję do faworyzowania ludzi, którzy są ekspertami. Dostarczają odpowiedzi. Odpowiedzi są ostateczne i zamknięte; pytania są otwarte. Ponadto dobre pytanie może prowadzić do większej liczby pytań. Dlatego zespoły zajmujące się analizą danych powinny polegać na kimś, kto koncentruje się wyłącznie na zadawaniu pytań. Ta osoba powinna rozumieć biznes, ale musi również wyjść poza biznes. Jednym z najlepszych nazwisk, jakie słyszałem w tej roli, jest kierownik badań. Ujmuje główną odpowiedzialność osoby za prowadzenie pytań i prowadzenie badań. Leadem badawczym powinien być ktoś ze strony biznesowej, który popycha zespół do zadawania ciekawych pytań. Powinna zacząć od pracy z zespołem, aby zadać pytania lub zidentyfikować kluczowe problemy. Może umieścić je na ścianie pytań lub uporządkować w karteczki samoprzylepne. Kierownik badania ma trzy obszary odpowiedzialności:

- Zidentyfikuj założenia
- Pytania motywujące
- Poznaj firmę

Te trzy obszary są ze sobą ściśle powiązane. Jako kierownik badań czasami musisz wyjść poza swoje doświadczenie i podejść do firmy tak, jakbyś widział ją po raz pierwszy. Wymaga to trochę wysiłku i musisz poświęcić się przyjęciu nowej perspektywy. Dobry lead badawczy będzie intuicyjnie wiedział,

kiedy pojawi się interesujące pytanie. Tak stało się z dr Jasonem Steffenem. Był astrofizykiem i częstym podróżnikiem. Znalazł się w długiej kolejce, czekając na wejście do samolotu. Większość ludzi po prostu akceptuje to jako część podróży. Dr Steffen rozumiał biznes latania na tyle, by zdać sobie sprawę, że czekanie pasażerów w długiej kolejce jest nieefektywne i kosztowne. Zadał proste pytanie: „Czy istnieje lepszy sposób na wejście na pokład samolotu?” Jego doświadczenie w nauce pomogło mu znaleźć rozwiązanie. Wyobraził sobie ludzi wsiadających równolegle, omijając co drugi rząd. To było o wiele bardziej efektywne niż obecna metoda ludzi ustawiających się pojedynczo. (Niestety, ta praktyka nie została zastosowana z różnych powodów.) Dr Steffen przeanalizował wszystkie trzy działania dobrego kierownika badań. Miał pewną wiedzę o biznesie. Wiedział o lataniu. Był w stanie zakwestionować swoje założenia. Miliony ludzi ustawili się przed nim w kolejce bez zastanowienia się nad procesem wejścia na pokład. W końcu wymyślił proste pytanie, czy istnieje lepszy sposób? Nie musisz być naukowcem, aby wymyślać ciekawe pytania. Prowadzący badania powinien mieć jakieś pojęcie o biznesie, ale ta osoba nie musi być ekspertem. Na przykład Edward Land wynalazł aparat natychmiastowy Polaroid, ponieważ zainspirowało go pytanie od jego trzyletniej córki. Kiedy byli na wakacjach w Nowym Meksyku, zrobił zdjęcie zwykłym aparatem. Jego córka po prostu zapytała: „Dlaczego musimy czekać na zdjęcie?” Jego córka zakwestionowała proste założenie. Oba te pytania zapoczątkowały interesujące kierunki poszukiwań. Dr Steffen wymyślił sposób, aby ludzie mogli wsiadać do samolotu nawet o 30% wydajniej, a Edward Land stworzył aparat natychmiastowy, jeden z najbardziej innowacyjnych i najpiękniejszych produktów technologicznych XX wieku. Jednym z najbardziej pomocnych elementów wiodącej roli badawczej jest oddzielenie pytań od danych. Nie ma nic z natury technicznego w zadawaniu dobrych pytań. Pamiętaj, córka Edwarda Landa miała zaledwie trzy lata. Istnieje również wiele korzyści z oddzielenia osoby kierującej pytaniami od osoby, która szuka możliwych odpowiedzi. Istnieje nieodłączny konflikt interesów. Jeśli masz tylko mały zestaw danych, możesz ograniczyć się do prostych pytań, na które możesz już odpowiedzieć za pomocą swoich danych. Dobry lead badawczy może sprawić, że ponownie zastanowisz się nad rodzajem gromadzonych danych. W końcu jest to o wiele bardziej wartościowe.

Przedstawianie danych analitykowi danych

Twój zespół potrzebuje dobrych analityków danych. Analityk danych jest odpowiedzialny za zrozumienie danych, pozyskiwanie i czyszczenie danych, a następnie wyświetlanie danych w prostych raportach. Powinni współpracować z prowadzącymi badania, aby sprawdzić, czy coś nie wyskakuje z raportów. Powinni także polecić metody statystyczne lub stworzyć wizualizacje danych. Kierownik badań i analityk danych będą pracować ramię w ramię, aby uzyskać szczegółowe informacje. Prowadzący badania koncentruje się na najlepszych pytaniach, które można zadać, podczas gdy analityk danych stara się dostarczyć najlepsze raporty. Istnieje wiele zamieszania wokół różnych tytułów stanowisk związanych z danymi. Są statystycy, analitycy statystyczni, analitycy danych, naukowcy danych, inżynierowie danych, a nawet matematycy. Możesz nawet dowolnie dodać „szef” lub „senior” do każdego z nich. Główny statystyk może być znacznie starszy niż starszy analityk danych. Powodem całego zamieszania jest to, że wszyscy ludzie w tych zawodach robią coś bardzo podobnego. W taki czy inny sposób wszyscy praktykują naukę uczenia się na podstawie danych; po prostu doszli do roli na bardzo różne sposoby. Statystyki istnieją od setek lat. Dyscyplina ewoluowała, ponieważ rządy musiały zrozumieć własne dane demograficzne i ekonomiczne. Ma długą i bogatą historię. Amerykańskie Stowarzyszenie Statystyczne (ASA) jest jednym z najstarszych stowarzyszeń zawodowych w kraju. Z kolei analitycy danych wywodzą się z informatyki. Uczą się wydobywać znaczenie z relacyjnych i NoSQL baz danych. Skupiają się na prezentowaniu i odkrywaniu interesujących fragmentów danych, które wspierają podejmowanie decyzji. Analitycy danych są postrzegani jako interdyscyplinarni. Są analitykami danych, ale tworzą też oprogramowanie, pracują z matematyką, znajdują się na biznesie i zadają ciekawe pytania. Jako były główny naukowiec w Bitly, Hilary

Mason stworzyła popularną definicję. Postrzega analityków danych jako analityków danych, którzy znają się również na matematyce, tworzeniu oprogramowania, inżynierii i hakowaniu. Mogą zrobić kolejny krok. Zamiast tworzyć raporty, mogą zacząć szukać informacji. Zespół analityków danych dzieli obowiązki, które zwykle spadają na jednego naukowca danych. Często prośenie jednej osoby o zrozumienie danych i biznesu oraz zadawanie ciekawych pytań to po prostu zbyt wiele. Dobry zespół zajmujący się analizą danych potrzebuje analityka danych, który również wie trochę o tworzeniu oprogramowania, a większość analityków danych już uważa, że wiedza o tworzeniu oprogramowania jest konieczna. Wiele świetnych narzędzi do wizualizacji wymaga kodowania oprogramowania. Python i R to dwa najpopularniejsze języki do eksploracji i wyświetlania danych. Analityk danych ma trzy główne obszary odpowiedzialności w zespole data science przygotowującym dane, wybierającym narzędzia, a następnie prezentującym wyniki.

Przygotowanie danych i wybór narzędzi

Przygotowanie danych i wybór narzędzi idą w parze. Musisz wybrać narzędzia do przygotowania danych. Dlatego jako analityk danych spędzasz większość czasu na przygotowywaniu danych. Musisz znaleźć najlepszy sposób na uzyskanie danych — za pośrednictwem internetowych interfejsów API, zbieranie ich ze stron lub zbieranie ich z różnych części organizacji — a następnie przeglądanie danych. Czyszczenie danych zwiększa ich przydatność, poprawiając różne pola lub dodając brakujące dane, takie jak rozwijanie skrótów lub poprawianie błędnie napisanych słów.

Prezentacja wyników

Jednym z głównych wyzwań dla analityka danych jest praca z prowadzącym badania w celu eksploracji danych, ustalenia, czy coś się wyróżnia, oraz tworzenia spostrzeżeń i raportów. Czasami bogactwo danych może prowadzić do niedostatku wglądu. Jeśli analityk danych przeciąży zespół, może to w rzeczywistości ograniczyć sposób, w jaki wszyscy interpretują informacje. Aby tego uniknąć, analityk danych musi ściśle współpracować z prowadzącym badaniem, aby wyjaśnić kompromisy w swoich raportach. Często w przypadku statystyk to, czego nie widzisz, jest kluczowe dla twojego zrozumienia. Czasami jest to celowe; innym razem tak nie jest. Powiedzmy, że kierownik badania chce zobaczyć zestawienie wszystkich mężczyzn i kobiet, którzy robią zakupy na stronie internetowej z butami do biegania i pokaż podział według grup wiekowych. Jako analityk danych możesz zapytać o przedziały wiekowe. Czy tworzysz wspornik co pięć lat czy co dziesięć lat? Jeśli tworzysz jeden na pięć lat, masz 18 lub 19 nawiasów, które są przekrzywione w środku, ponieważ prawdopodobnie jest mniej biegaczy w wieku poniżej 18 lat lub starszych niż 90 lat. Raport może być również trudny do odczytania. Nie miałbyś takiej szczegółowości, jakiej możesz potrzebować w przypadku osób w wieku od 20 do 40 lat. Może być duża różnica między 35- i 40-latkami. Jeśli przekroczysz ten przedział wiekowy, możesz fałszywie przedstawiać dane. Może się wydawać, że jest mniej osób w wieku od 30 do 40 lat. Dlatego ważne jest, aby analityk danych był przejrzysty w sposobie prezentowania danych. Są tu decyzje, które wpływają na fabułę. Cały zespół powinien pracować, aby być na bieżąco z tymi decyzjami i przekazać dokładną historię. Całe to podejmowanie decyzji normalnie leżałoby w gestii analityka danych. Jednak praca z badaniami prowadzi do podjęcia tych decyzji, co pomaga całemu zespołowi lepiej zrozumieć dane.

Bądź na bieżąco z kierownikiem projektu

Jedną z najtrudniejszych części pracy w zespole ds. analityki danych jest umożliwienie wykonywania spostrzeżeń. Praca powinna zacząć przypominać trwający eksperyment naukowy. Dostajesz trochę danych. Prowadzenie badań prowadzi do interesujących pytań. Pytania generują więcej raportów, co zwykle prowadzi do większej liczby pytań. W tym wirze eksploracji zespół musi skoncentrować się na celu zebrania danych. Potrzebny jest również mechanizm wyjaśniający spostrzeżenia reszcie

organizacji. Niestety wiąże się to z wieloma spotkaniami, które mogą być stratą czasu rzeczywistego. Jest to szczególnie ważne, gdy praca jest twórcza. Kiedyś pracowałem dla organizacji, która miała bardzo kreatywny zespół ds. data science. Wymyślali nowe i interesujące sposoby wykorzystania ogromnych danych kart kredytowych firmy. Przez pierwsze kilka miesięcy zespół zajmujący się analizą danych był w większości pozostawiony sam sobie z badaniem danych. Ostatecznie zespół przedstawił kilka naprawdę interesujących spostrzeżeń. W miarę jak ich spostrzeżenia stawały się coraz bardziej interesujące, reszta organizacji stawała się bardziej ciekawa. Departamenty zaczęły wzywać członków zespołu do wygłaszania prezentacji. Te spotkania jeszcze bardziej zainteresowały pozostałe zespoły. Doprowadziło to do większej liczby spotkań. Po kilku miesiącach niektórzy członkowie zespołu zajmującego się analizą danych spotykali się na spotkaniach do 20 godzin tygodniowo. Przeszli od praktyków ds. data science do prezydentów. Brak tych kluczowych osób z zespołu analityków danych spowodował resztę zespołu do tego stopnia, że uzyskiwali mniej spostrzeżeń. Spędzili znacznie mniej czasu na przeglądaniu swoich danych. Te same działy, które prosiły o spotkania, zaczęły pytać, dlaczego zespół ds. analityki danych nie znajduje tylu spostrzeżeń. To paradoks w wielu organizacjach. Im bardziej interesująca jest Twoja praca dla reszty organizacji, tym więcej czasu spędzasz na rozmowach o swojej pracy na spotkaniach, co oznacza mniej czasu na pracę. Najlepszym sposobem na przerwanie tego cyklu jest posiadanie skutecznego kierownika projektu. Menedżerowie projektów są bardzo dobrzy w ochronie zespołu ds. analityki danych przed zejściem z kursu. Mogą to zrobić, reprezentując zespół na spotkaniach. Kierownik projektu w zespole ds. data science jest odpowiedzialny za trzy główne obszary: demokratyzację danych (udostępnienie ich szeroko), dzielenie się wynikami i egzekwowanie organizacyjnego uczenia się. Kierownik projektu działa zarówno jako tarcza, jak i buldożer; tarcza chroniąca zespół przed zbyt dużą liczbą spotkań, co pomoże utrzymać zespół w pracy, oraz buldożer do przełamywania barier i uzyskiwania dostępu do ukrytych danych.

Demokratyzacja danych

Kiedy pracujesz w dużej organizacji, dostęp do wszystkich danych staje się coraz trudniejszy. Kierownik projektu może pomóc przebić się przez te silosy, aby cała organizacja mogła na tym skorzystać. Załóżmy na przykład, że zespół zajmujący się analizą danych chce uzyskać dostęp do danych znajdujących się na jakimś zakurczonym serwerze w rogu centrum danych. Kiedy zespół próbuje uzyskać do niego dostęp, dowiaduje się, że jest on niedostępny i dostępny tylko dla określonego działu. Twój kierownik projektu jest odpowiedzialny za próbę przekonania tego działu do udzielenia zespołowi ds. analizy danych dostępu do tych danych.

Udostępnianie wyników

Kierownicy projektów pracują również nad dystrybucją informacji. To oni chodzą na spotkania i prezentują wyniki zespołu. Jeśli ktoś z innego działu chce uzyskać dostęp do danych zespołu, kierownik projektu pracuje nad przyznaniem mu dostępu. Demokratyzacja danych i dystrybucja wyników są ze sobą ściśle powiązane. Jest to droga dwukierunkowa: jeden sposób zapewnia dostęp do danych organizacyjnych, a drugi zapewnia dostęp do spostrzeżeń zespołu reszcie organizacji. Każdy z nich ma swój własny zestaw spotkań i wyzwań.

Wymuszanie uczenia się organizacyjnego

Ostatnim obszarem dla kierownika projektu jest wymuszanie uczenia się - wyciąganie spostrzeżeń i sprawianie, by były one wykonalne. Pod koniec dnia zespół nadal będzie oceniany na podstawie tego, czego nauczy się organizacja. Ktoś musi śledzić i przekształcać spostrzeżenia w produkty lub zmiany. Wyobraź sobie, że organizacja dysponująca danymi karty kredytowej znalazła coś interesującego, na przykład niewielki wzrost danych transakcyjnych w porze lunchu. Kierownik projektu zamieniłby ten wgląd w coś wykonalnego. Może również współpracować z grupą ds. infrastruktury, aby w tym czasie

skalować swoją technologię lub współpracować z grupą marketingową, aby tworzyć promocje w porze lunchu. Wymuszanie tej nauki zajmuje dużo czasu i dużo spotkań. Dobry kierownik projektu utrzymuje zespół na dobrej drodze, biorąc na siebie tę odpowiedzialność.

Praca w zespole

Po utworzeniu zespołu cały zespół pracuje razem, aby sprawdzić, czy mogą stworzyć interesujący model danych, który pokazuje trendy w danych. Być może zauważysz silną korelację między kilkoma różnymi elementami, a następnie spróbujesz wspólnie odgadnąć, co powoduje korelację. Załóżmy, że Twój zespół pracuje dla magazynu internetowego. Na górze każdej historii znajduje się link, który pozwala czytelnikom udostępnić artykuł w swojej sieci społecznościowej. Analityk danych w Twoim zespole ocenia historie, które były najczęściej udostępniane przez czytelników, i przygotowuje raport dla zespołu, dzięki czemu możesz omówić wyniki. Z kolei prowadzący badanie może zapytać: „Co sprawia, że te artykuły są tak popularne? Czy są najczęściej czytane czy najczęściej udostępniane? Czy istnieją pewne tematy, które zwiększają prawdopodobieństwo udostępnienia artykułu? Czy są frazy kluczowe?” Twój zespół pracuje razem, aby stworzyć model, który pokazuje korelację między niektórymi tematami a prawdopodobieństwem udostępnienia tego tematu. Może zespół tworzy własną identyfikację tematu. Prowadzenie badań jest tutaj kluczowe, ponieważ to ona wie najwięcej o biznesie. Może wiedzieć wystarczająco dużo o czytelnikach, aby odgadnąć pewne kategorie tematyczne. Może być najlepszym źródłem do wymyślania słów kluczowych, takich jak „zajawka”, „niewyraźne zdjęcie”, „dziwaczny humor” lub „plotki”, które sugerują plotki technologiczne. (Ktoś z doświadczeniem wydawniczym miałby wiele przydatnych sugestii). Zespół mógłby nawet stworzyć pierwszą wersję programu, który spróbuje sortować artykuły do tych kategorii na podstawie słów kluczowych. Analityk danych tworzy raporty i rozwija aplikację, która identyfikuje temat historii. Następnie tworzy model danych, który dokładnie sortuje najczęściej udostępniane artykuły. (Ta osoba może nie wiedzieć zbyt wiele o publikowaniu, ale wie, jak przetwarzać dane.) Wreszcie zespół korzysta z analiz predykcyjnych, aby zastosować ten model w przyszłości. Masz teraz aplikację, która może dokładnie przewidzieć, kiedy nowy artykuł prawdopodobnie zostanie udostępniony przez wielu czytelników.

W tym miejscu wkracza kierownik projektu, przejmuje ten nowy wgląd i czyni go praktycznym. Przekazuje wyniki innym zespołom i współpracuje z kierownictwem w celu doskonalenia organizacji. Może nawet współpracować z działem marketingu, aby umieścić najbardziej dochodowe reklamy w artykułach, które z większym prawdopodobieństwem będą chętnie udostępniane przez czytelników. Częścią sposobu myślenia na temat nauki o danych jest uznanie, że zachęcanie do innowacji wymaga współpracy kierownika badań i analityków danych w celu promowania nowych pomysłów. Prowadzenie badań musi prowadzić do ciekawych pytań. Muszą także wspierać różnorodność opinii. Mogą chcieć sprowadzić ludzi z innych części organizacji. Jeśli prowadzisz swój zespół wyłącznie z analitykami danych, prawdopodobnie brakuje Ci znacznej różnorodności opinii. Będzie zbyt wiele podobieństw między ich wykształceniem a środowiskiem. Bardziej prawdopodobne jest, że szybko dojdą do konsensusu i zaśpiewają w monotonnym refrenie. Kiedyś pracowałem ze szkołą podyplomową, która próbowała zwiększyć wskaźnik ukończenia studiów, patrząc na dane z przeszłości. Okazało się, że najlepszy pomysł wyszedł od kierownika projektu, który był również zapalonym pletwonurkiem. Spojrzał na dane demograficzne i zasugerował, że system znajomych może zwiększyć liczbę uczniów, którzy przeszli przez cały program. To była powszechna praktyka podczas treningu nurkowego. Nikt nie mógł zaplanować jego wglądu. Po prostu pochodziło z jego życiowego doświadczenia. Ta forma kreatywnego odkrywania jest znacznie bardziej powszechna, niż większość organizacji zdaje sobie sprawę. W rzeczywistości raport urzędu patentowego sugerował, że prawie

połowa wszystkich odkryć jest wynikiem prostego zbiegu okoliczności. Zespół szukał rozwiązania problemu, a czyjaś wiedza lub doświadczenie poprowadziły ich w zupełnie nowym kierunku.

PODSUMOWANIE

Dowiedziałeś się, że jeden z najskuteczniejszych modeli zespołowych w nauce o danych obejmuje następujące role: kierownik badań, analityk danych i kierownik projektu. Prowadzenie badań ma trzy obszary odpowiedzialności: znajdowanie założeń, kierowanie pytaniami i poznawanie biznesu. Analityk danych przygotowuje dane, dobiera narzędzia i prezentuje wyniki. Na koniec kierownik projektu zbiera potrzebne dane z różnych działów, a następnie dzieli się wynikami zespołu i wymusza uczenie się organizacji.