Dokonywanie skutecznej analizy

Ta część została napisana w kontekście naukowców zajmujących się danymi, którzy skupiają się na nauce o decyzjach i analityce – ludziach, którzy wykorzystują dane do dostarczania pomysłów i sugestii dla biznesu. Chociaż inżynierowie zajmujący się uczeniem maszynowym muszą również przeprowadzać analizę przed budowaniem i wdrażaniem modeli, niektóre treści dotyczące zarządzania interesariuszami z ładnymi wizualizacjami są mniej istotne. Jeśli jesteś inżynierem zajmującym się uczeniem maszynowym i czytasz to, nie martw się; jest to nadal dla Ciebie bardzo istotne, a pokochasz Część 11, która obejmuje wdrażanie modeli do produkcji. Podstawą wielu zawodów związanych z analityką danych jest dokonywanie analiz: krótkich dokumentów, które wykorzystują dane do wyjaśnienia sytuacji biznesowej lub rozwiązania problemu biznesowego. Nowoczesne firmy zbudowane są na raportowaniu i analizach. Ludzie, którzy podejmują decyzje, nie czują się komfortowo bez danych, które uzasadniają ich wybory, a analitycy danych to jedni z najlepszych ludzi, którzy potrafią znaleźć znaczenie danych. Analizy są również ważne przy tworzeniu narzędzi uczenia maszynowego, ponieważ zanim będzie można zbudować model uczenia maszynowego, należy zrozumieć kontekst zestawu danych. Stworzenie analizy, która potrafi zebrać ogromną ilość danych firmowych i przekształcić je w zwięzły wynik, który wyjaśnia sprawę, jest niezwykle trudne i praktycznie sztuką. Jak można oczekiwać, że dana osoba weźmie tabele z milionami rekordów informacji historycznych, z których każda zawiera złożoność i niuanse, i zamieni je w ostateczne „Tak, dane mówią, że ten pomysł jest dobry”? Odkrywanie, co ma sens matematycznie, czym interesuje się biznes i jak wypełnić lukę między nimi, nie jest czymś, co powinieneś wiedzieć, jak to zrobić w naturalny sposób . Omówimy tu podstawy budowania analizy, aby zrozumieć, jak dostarczać firmie sensowne analizy. Korzystając z umiejętności tu opisanych , powinieneś być w stanie szybciej rozwijać się w swojej karierze data science. Czym tak naprawdę jest analiza? Analiza to zazwyczaj prezentacja programu PowerPoint, plik PDF lub Word albo arkusz kalkulacyjny programu Excel, który można udostępnić naukowcom nie zajmującym się danymi, zawierający spostrzeżenia z danych i wyświetlające je wizualizacje. Wykonanie analizy zajmuje zazwyczaj od jednego do czterech tygodni, przy czym analityk danych musi zebrać dane, uruchomić na nim kod dla metod statystycznych i uzyskać ostateczny wynik. Po zakończeniu kod nie jest zmieniany, dopóki analiza nie musi zostać ponownie uruchomiona kilka miesięcy później, a być może nigdy. Przykładowe analizy obejmują

* Analizowanie danych z ankiet klientów, aby zobaczyć, które produkty cieszą się największą satysfakcją

* Patrząc na dane dotyczące lokalizacji, z których składane są zamówienia, aby wybrać lokalizację nowej fabryki

* Wykorzystanie historycznych danych z branży lotniczej do przewidzenia, które miasta będą potrzebować więcej tras do nich

Te przykłady mają różne poziomy złożoności technicznej; niektóre wymagają jedynie podsumowania i wizualizacji danych, podczas gdy inne potrzebują metod optymalizacji lub modeli uczenia maszynowego, ale wszystkie z nich odpowiadają na jednorazowe pytanie.

Raportowanie a analiza

Raport i analiza są podobne, ale nie takie same. Raport to coś, co jest generowane cyklicznie bez większych zmian strukturalnych między wersjami. Na przykład miesięczny raport finansowy może być dużym arkuszem kalkulacyjnym programu Excel, który co miesiąc aktualizuje się o nowe liczby. Celem raportu jest uświadomienie ludziom, jak zmieniają się dane. Analiza to coś, co jest wykonywane jednorazowo, aby odpowiedzieć na głębsze pytanie. Analiza pozyskania klientów może być wykonana

w R na temat tego, jak nowi klienci kupują produkty, a wyniki zostaną umieszczone w prezentacji PowerPoint. Raporty są zwykle wypełnione liczbami i metrykami, podczas gdy analizy skupiają się na dostarczeniu jednego głównego wyniku. Większość cech dobrej analizy dotyczy dobrego raportu, więc używamy analizy w znaczeniu obu, chyba że wyraźnie zaznaczono inaczej. Więc co sprawia, że analiza jest dobra? Dobra analiza ma pięć cech:

* Odpowiada na pytanie. Analiza zaczyna się, gdy ktoś zadaje pytanie, więc aby analiza miała sens, musi udzielić odpowiedzi. Jeśli zaproponowane pytanie brzmiało „Która z tych dwóch stron powoduje, że więcej klientów kupuje produkty?", analiza powinna wykazać, która strona powoduje większą sprzedaż. Ta odpowiedź może nawet brzmieć „Nie mamy wystarczającej ilości informacji do powiedzenia", ale musi to być bezpośrednia odpowiedź na pytanie.

* Robi się szybko. Odpowiedzi na pytania biznesowe wpłyną na decyzje, które mają terminy. Jeśli analiza trwa zbyt długo, decyzja zostanie podjęta bez analizy. Powszechnie oczekuje się, że analiza zostanie ukończona w ciągu miesiąca.

* Można ją udostępniać. Analiza musi zostać udostępniona nie tylko osobie, która poprosiła o jej wykonanie, ale także komuś , z kim ta osoba chce się nią podzielić. Jeśli analiza obejmuje na przykład fabułę, fabuła nie może po prostu żyć w skrypcie R lub Python; musi być w formacie zrozumiałym dla ludzi, takim jak PowerPoint.

* Jest samowystarczalna. Ponieważ nie można przewidzieć, kto zobaczy analizę, sama musi być zrozumiała. Wykresy i tabele muszą mieć jasne opisy, osie powinny być oznakowane, wyjaśnienia w analizie powinny być spisane, a analiza powinna w miarę możliwości unikać odwoływania się do innych prac.

* Można go ponownie odwiedzić. Większość pytań zostanie zadanych ponownie w przyszłości. Czasami odpowiedź na nie oznacza ponowne wykonanie dokładnie tej samej pracy, na przykład ponowne uruchomienie grupowania. Czasami trzeba zastosować podejście gdzie indziej, na przykład zmienić dane wejściowe z klientów europejskich na klientów azjatyckich.

Cechy te łączą się w ogólny temat „Dobra analiza to coś, co pomaga naukowcom, którzy nie zajmują się danymi". Reszta tego rozdziału ma strukturę chronologiczną obejmującą etapy analizy, począwszy od wstępnego wniosku o analizę, a skończywszy na przekazaniu raportów. Chociaż nie każda analiza będzie przebiegać zgodnie z tymi krokami, większość będzie (lub powinna). Gdy zaznajomisz się z wykonywaniem analiz, możesz mieć ochotę pominąć niektóre kroki, ale te skróty są właśnie działaniami, które powodują, że starsi specjaliści ds. danych popełniają błędy.

Analizy dla różnych typów analityków danych

W zależności od Twojej roli jako analityka danych sytuacje, w których będziesz przeprowadzać analizy, będą się znacznie różnić:

* Naukowiec decyzyjny - dla tego typu naukowców zajmujących się danymi, wykonywanie analiz jest podstawową funkcją pracy. Naukowcy zajmujący się decyzjami nieustannie zagłębiają się w dane, aby odpowiedzieć na pytania, a te pytania należy przekazać firmie. Kluczowym narzędziem do tego jest analiza.

* Inżynier ds. uczenia maszynowego - chociaż inżynier ds. uczenia maszynowego koncentruje się na tworzeniu i wdrażaniu modeli, analizy są nadal przydatnymi narzędziami do dzielenia się wydajnością modeli. Analizy służą do pokazania wartości w budowaniu nowego modelu lub tego, jak modele zmieniają się w czasie.

* Analitycy-analitycy, którzy są analitykami danych, którzy skupiają się głównie na metrykach i wskaźnikach KPI dla firmy, zwykle tworzą wiele raportów. Tworzą strumień powtarzających się danych dla firmy, często w Excelu, SQL, R lub Pythonie. Chociaż ci eksperci od analityki będą dokonywać analiz, muszą myśleć o łatwości utrzymania pracy bardziej niż o innych rolach, ponieważ muszą to powtarzać tak często.

Prośba

Analiza zaczyna się od prośby o odpowiedź na pytanie biznesowe. Osoba z innej części firmy lub Twój menedżer przyjdzie do Ciebie z pytaniem typu „Czy możesz sprawdzić, dlaczego sprzedaż gadżetów w Europie była niska w grudniu?" lub „Czy nasi klienci z małych firm zachowują się inaczej niż nasi więksi?" W zależności od poziomu wiedzy technicznej osoby pytającej, możesz otrzymać nieprecyzyjne zapytanie („Dlaczego sprzedaż spada?") lub precyzyjne („Jakie atrybuty są skorelowane z niższą średnią wartością zamówienia?"). Analiza jest tworzona wokół pytania biznesowego, ale nie można wykonywać analizy danych w przypadku pytania biznesowego. Pytania związane z nauką o danych to np. „Jak pogrupować te punkty danych?" oraz „Jak prognozujemy sprzedaż?" Naukowiec zajmujący się danymi musi wykonać zadanie przekształcenia tego pytania biznesowego w pytanie dotyczące analizy danych, udzielenia odpowiedzi na pytanie z zakresu analizy danych i zwrócenia odpowiedzi biznesowej. Ta praca jest trudna. Zrozumienie, w jaki sposób pytania związane z nauką o danych i pytania biznesowe są powiązane, wymaga połączenia doświadczenia z rodzajem problemu oraz zrozumienia, w jaki sposób wyniki różnych metod statystycznych mogą być potencjalnie użyteczne. Przepływ pracy od pytań biznesowych do pytań i odpowiedzi związanych z nauką o danych i wreszcie z powrotem do odpowiedzi biznesowej został opracowany przez Renee Teate. Pytanie biznesowe pochodzi od interesariuszy, którzy chcą wiedzieć, jak kierować marketing do różnych klientów. Analityk danych musi dowiedzieć się, co to żądanie oznacza w kategoriach matematycznych - w tym przykładzie grupowanie danych klientów. Po zakończeniu procesu specjalista ds. danych ma odpowiedź w zakresie nauki o danych (na przykład zestaw trzech grup zgrupowanych punktów danych). Wreszcie, analityk danych musi przekonwertować tę odpowiedź z powrotem na coś, co firma by zrozumiała, na przykład grupy takie jak „nowi klienci" lub „wiele wydający". Zanim zaczniesz przyglądać się danym i pisać kod w celu rozwiązania pytania z zakresu nauki o danych, musisz wykonać podstawową pracę, aby jak najlepiej zrozumieć pytanie biznesowe. Musisz zrozumieć, jaki jest kontekst analizy, aby jak najlepiej dostarczyć coś przydatnego. Kto prosi o przeprowadzenie analizy i jaki jest ich stosunek do ich drużyny? Jaki jest ich motyw? Czy mają bardzo konkretne pytanie, na które chcą uzyskać odpowiedź, lub niejasne, ogólne pojęcie problemu i nadzieję, że dane mogą się do tego przydać? Czy wydaje się, że możesz mieć dane, aby rozwiązać ten problem? Jeśli nie, co trzeba by go zdobyć? Zadawanie pytań nie tylko pomaga zrozumieć, jak rozwiązać problem, ale także pomaga zrozumieć, do czego będzie on używany. Wielu analityków danych spędziło tygodnie na analizach tylko po to, by odkryć, że nie było takiej potrzeby, ponieważ interesariusz był „po prostu ciekawy". Odpowiedzi na te pytania są zazwyczaj udzielane podczas 30-60-minutowego spotkania inauguracyjnego z osobą składającą wniosek oraz wszystkimi innymi osobami zaangażowanymi w pracę. Jako osoba przeprowadzająca analizę możesz nie organizować spotkania, ale jeśli nie masz go w swoim kalendarzu, warto zaplanować spotkanie. Jeśli nie spotkałeś wcześniej osoby, która zleca analizę, to spotkanie jest dobrym momentem na przedstawienie się i zapoznanie się z jej pracą. Hipotetyczny przykładowy zestaw podstawowej wiedzy wyglądałby mniej więcej tak:

* Kto prosi o analizę? Poprosiła o to Julia z zespołu ds. produktu widget.

* Jaki jest motyw? Sprzedaż widżetów spadła w tym miesiącu o 10 procent, a zespół biznesowy nie wie, dlaczego.

* Jaka jest prośba? Zespół chce wykorzystać dane, aby sprawdzić, czy spadek sprzedaży widżetów dotyczył jednej części kraju.

* Jaka decyzja zostanie podjęta? Decyzja dotyczy tego, czy produkt widget powinien zostać wycofany, czy nie.

* Czy posiadamy wymagane dane? Tak, do analizy potrzebne są zamówienia klientów poprzez wysyłkę kodu pocztowego, który jest dostępny w bazie zamówień.

Wiedza, czy masz dane, które mogłyby wiarygodnie odpowiedzieć na pytanie, jest naprawdę ważna. Ostatnią rzeczą, którą chcesz zrobić, to spędzić kilka tygodni na pracy nad analizą, tylko po to, aby wrócić do interesariuszy bez czegokolwiek, z czego mogliby skorzystać. Przykładem sytuacji, w której nie miałbyś danych, byłoby coś takiego: w firmie detalicznej interesariusz chce wiedzieć, ile zamówień złożył każdy klient, ale ponieważ klienci płacą gotówką, nie ma możliwości wykorzystania istniejące dane, aby powiedzieć, kto złożył każde zamówienie. W takiej sytuacji najlepiej jest być szczerym wobec wszystkich zaangażowanych osób i dać im do zrozumienia, że to, o co proszą, nie jest możliwe. Inne osoby mogą zaproponować alternatywne sposoby wykorzystania danych, które mogą być wystarczająco zbliżone do tego, na co liczyłeś, lub być może będziesz musiał wyjaśnić, dlaczego alternatywy również nie będą działać. Jeśli to w ogóle możliwe, zaproponuj plan, który może kiedyś uzyskać wymagane dane. W poprzednim przykładzie program lojalnościowy umożliwiłby powiązanie zamówień z konkretnym klientem, a tym samym naprawienie problemu z danymi, chociaż utworzenie tego programu zajęłoby trochę czasu. Pozostałe pytania, takie jak to, kim jest dana osoba i dlaczego składa wniosek, są przydatne do tworzenia planu analizy.

Plan analizy

Dla naukowców zajmujących się danymi nie ma nic przyjemniejszego niż zagłębienie się w niektóre dane, aby odpowiedzieć na pytania. Załadujmy dane! Pogrupuj to! Podsumuj to! Dopasuj model i wykreśl wyniki! Niestety, ponieważ istnieje nieskończona liczba sposobów podsumowywania i modelowania danych, możesz spędzić tygodnie pracując z danymi tylko po to, aby odkryć, że nic, co udało się wygenerować, nie odpowiada na zaproponowane pytanie biznesowe. Najgorsze jest uświadomienie sobie, że nie zrobiłeś czegoś istotnego. Zdarza się to często badaczom danych, zwłaszcza młodszym, którzy nie zostali „poparzeni" zbyt wiele razy. Jednym z rozwiązań tego problemu jest posiadanie poręczy zabezpieczającej, która utrzyma Cię na właściwej drodze i wykona odpowiednią pracę. Plan analizy jest tą barierą ochronną. Pomysł polega na tym, że zanim zaczniesz patrzeć na dane, zapisujesz wszystko, co planujesz zrobić z danymi. Następnie, w miarę postępu analizy, śledzisz, ile z planu udało Ci się zrealizować. Kiedy zrobisz wszystko zgodnie z planem, gotowe! Masz nie tylko sposób na sprawdzenie, czy nie masz planu, ale także narzędzie do śledzenia postępów i rozliczania się. Możesz nawet użyć go na spotkaniach z przełożonym, aby omówić, jak się sprawy mają. Tworząc plan analizy, chcesz, aby praca w planie była wykonalna. „Stwórz regresję liniową sprzedaży według regionu" to coś, co możesz napisać w kodzie, podczas gdy „Dowiedz się, dlaczego sprzedaż spadła" nie jest czymś, co możesz po prostu zrobić; to wynik robienia innych rzeczy. Jeśli zadania w planie są wykonalne, łatwo będzie stwierdzić, czy robisz postępy. Ułatwi to również przeprowadzenie analizy, ponieważ nie będziesz musiał martwić się o to, co dalej. Zamiast tego będziesz mógł spojrzeć na plan analizy i wybrać następne zadanie do wykonania. Do przygotowania kilku pierwszych planów analizy zdecydowanie zalecamy skorzystanie z następującego szablonu:

* U góry - Podaj tytuł analizy, kim jesteś (w przypadku, gdy analiza będzie udostępniana innym) oraz cel analizy.

* Sekcje - Każda sekcja powinna być ogólnym tematem analizy. Praca analityczna wykonana w ramach każdej sekcji powinna być niezależna (nie polegać na pracy innych sekcji), tak więc każda sekcja powinna być wykonywana przez inną osobę. Każda sekcja powinna mieć listę zadań.

* Pierwszy poziom list sekcji - Pierwszym poziomem list sekcji powinno być każde zadane pytanie. Ta sekcja pomoże każdemu zapamiętać, dlaczego wykonujesz tę konkretną pracę, a jeśli wszystkie pytania zostaną udzielone pomyślnie, temat głównej sekcji należy uznać za zrozumiały.

* Drugi poziom list sekcji - Drugi poziom list powinien zawierać faktyczne zadania do wykonania, które można odhaczać w trakcie wykonywania pracy. Mogą to być na przykład rodzaje modeli do uruchomienia, a opisy powinny być na tyle szczegółowe, aby w dowolnym momencie można było konkretnie stwierdzić, czy praca została zakończona.

Kiedy tworzysz plan analizy, podziel się nim z przełożonym i interesariuszem składającym wniosek. Powinni albo dać sugestie, jak ją ulepszyć, albo zatwierdzić pracę. Zatwierdzony plan analizy stanowi uzgodnioną podstawę pracy. Jeśli po przeprowadzeniu analizy interesariusz zapyta, dlaczego zrobiłeś coś w ten sposób, możesz odwołać się do planu analizy i pierwotnych celów. Prawdopodobnie podczas przeprowadzania analizy zdasz sobie sprawę, że pominąłeś coś ważnego z planu analizy lub masz nowy pomysł, którego wcześniej nie brałeś pod uwagę. To całkowicie w porządku; po prostu zaktualizuj plan i poinformuj interesariuszy, że wprowadzasz zmianę. Ponieważ masz ograniczenia czasowe, być może będziesz musiał usunąć mniej ważne zadanie z istniejącego planu. Ale znowu, plan analizy jest przydatny, ponieważ tworzy rozmowę na temat tego, co usunąć, zamiast zmuszać cię do wykonania niemożliwej ilości pracy.

Robienie analizy

Po napisaniu planu analizy możesz rozpocząć samą analizę! Praca zaczyna się od importu danych, aby można było nimi manipulować i czyścić. Następnie wielokrotnie przekształcasz dane, podsumowując, agregując, modyfikując, wizualizując i modelując je. Gdy dane są gotowe, przekazujesz tę pracę innym. W kolejnych sekcjach pokrótce omówimy niektóre kwestie, o których należy pamiętać podczas przeprowadzania analizy w środowisku pracy. Całe książki poświęcone temu tematowi mogą również nauczyć Cię kodu do przeprowadzenia analizy w wybranym przez Ciebie języku.

Importowanie i czyszczenie danych

Zanim będziesz mógł zająć się pytaniami w planie analizy, musisz mieć dane w miejscu, w którym możesz nimi manipulować i w formacie, którego możesz użyć. Zwykle oznacza to możliwość załadowania go w R lub Pythonie, ale może obejmować użycie SQL lub innych języków. Niemal zawsze to zadanie zajmie Ci więcej czasu, niż się spodziewasz. W tym procesie może pojawić się wiele niespodzianek. Kilka z tych wielu horrorów to

* Problemy z łączeniem się z firmowymi bazami danych w konkretnym zintegrowanym środowisku programistycznym (IDE)

* Problemy z nieprawidłowymi typami danych (takimi jak liczby jako ciągi)

* Problemy z dziwnymi formatami czasu („rok-dzień-miesiąc” zamiast „rok-miesiąc-dzień”)

* Dane wymagające formatowania (być może każdy identyfikator zamówienia zaczynał się od „ID-” i musisz to usunąć)

* Rekordy, których brakuje w danych

Co gorsza, żadna z tych prac nie wygląda na produktywną dla osób nietechnicznych; nie możesz pokazać interesariuszom przekonującego wykresu pokazującego, jak działa sterownik bazy danych, ani nie zrozumieją, że manipulacja ciągami pomaga im w rozwiązaniu ich problemów biznesowych. Tak żmudne, jak to zadanie, chcesz szybko przejść do eksploracji danych. Kiedy pracujesz nad importowaniem i porządkowaniem danych, weź pod uwagę, że masz podwójne zadanie: poświęcaj jak najmniej czasu na wszystko, co nie będzie potrzebne, i jak najwięcej czasu na pracę, która pomoże w dalszej pracy. Jeśli masz kolumnę z datami, które są przechowywane jako ciągi i masz wątpliwości, czy kiedykolwiek będziesz potrzebować tej kolumny, nie trać czasu na zmianę ciągów na właściwy format daty i godziny. Z drugiej strony, jeśli uważasz, że potrzebujesz tej kolumny, zdecydowanie wykonaj pracę tak szybko, jak to możliwe, ponieważ chcesz mieć czysty zestaw danych do analizy. Trudno powiedzieć z wyprzedzeniem, co będzie przydatne, ale jeśli spędzasz na czymś dużo czasu, zadaj sobie pytanie, czy naprawdę tego potrzebujesz. Podczas importowania i porządkowania danych możesz utknąć na wiele dni w jednym problemie, takim jak połączenie z bazą danych. Jeśli znajdziesz się w takiej sytuacji, masz do wyboru trzy opcje: (1) poprosić o pomoc, (2) znaleźć sposób na rozwiązanie problemu, którego można całkowicie uniknąć, lub (3) nadal próbować samodzielnie rozwiązać problem. Opcja (1) jest świetna, jeśli możesz to zrobić: starsza osoba może znaleźć szybkie rozwiązanie, a ty możesz uczyć się z tego, co zrobiła. Opcja (2) też jest świetna; zrobienie czegoś takiego jak użycie płaskiego pliku .csv zamiast połączenia z bazą danych pozwoli Ci przejść do analizy, która zapewnia wartość biznesową. Opcja (3) – próbuj i próbuj – jest czymś, czego powinieneś unikać za wszelką cenę. Jeśli spędzasz dni i dni nad jednym problemem, będziesz wyglądać, jakbyś nie był w stanie dostarczyć wartości. Jeśli coś jest dla ciebie nie do pokonania, porozmawiaj ze swoim przełożonym, co zrobić; nie tylko próbuj i miej nadzieję, że jakoś problem sam się rozwiąże. Po załadowaniu danych i sformatowaniu ich możesz zacząć z nich korzystać i znaleźć dziwne dane. Dziwne dane to wszystko, co wykracza poza podstawowe założenia. Gdybyś na przykład spojrzał na historyczne dane dotyczące lotów linii lotniczych i znalazł kilka lotów, które lądowały przed startem, byłoby to dziwne, ponieważ generalnie samoloty startują jako pierwsze! Inną dziwnością może być wszystko, od sklepu sprzedającego przedmioty, które mają ujemną cenę, po dane produkcyjne pokazujące, że jedna fabryka wyprodukowała tysiąc razy więcej przedmiotów niż fabryka podobna. Tego rodzaju dziwne artefakty pojawiają się cały czas w danych ze świata rzeczywistego i nie ma możliwości ich przewidzenia, dopóki sam nie spojrzysz na dane. Jeśli znajdziesz się w sytuacji, w której masz dziwne dane, nie ignoruj ich! Najgorsze, co możesz zrobić, to założyć, że dane są w porządku, a następnie, po tygodniach pracy nad analizą, dowiedzieć się, że dane nie były w porządku i zmarnować swoją pracę. Zamiast tego porozmawiaj ze swoim interesariuszem lub osobą odpowiedzialną za dane, których używasz, i zapytaj, czy są świadomi dziwności. W wielu przypadkach już o tym wiedzą i sugerują, abyś to zignorował. W przykładzie danych linii lotniczych możesz po prostu usunąć dane dotyczące lotów, które wylądowały przed startem. Jeśli okaże się, że dziwaczność była nieznana i mogłaby zagrozić analizie, musisz zbadać sposoby jej uratowania. Jeśli zamierzasz przeprowadzić analizę porównującą przychody i koszty, a co dziwne, w połowie Twoich danych brakuje kosztów, musisz sprawdzić, czy możesz pracować z samymi istniejącymi kosztami, czy z samymi przychodami. W pewnym sensie to podejście staje się analizą w analizie; robisz mini-analizę, aby sprawdzić, czy pierwotna analiza jest w ogóle wykonalna.

Eksploracja i modelowanie danych

Podczas części analizy dotyczącej eksploracji danych i modelowania przechodzisz przez plan analizy punkt po punkcie i próbujesz dokończyć pracę. Poniższe sekcje zawierają ogólne ramy dotyczące każdego punktu.

UŻYWAJ OGÓLNEGO PODSUMOWANIA I PRZEKSZTAŁCENIA

Zdecydowaną większość prac analitycznych można wykonać, podsumowując i przekształcając dane. Pytania takie jak „Ilu klientów mieliśmy każdego miesiąca?” można odpowiedzieć, pobierając dane klientów, grupując je na poziomie miesiąca, a następnie zliczając odrębną liczbę klientów w każdym miesiącu. Ta technika nie wymaga metod statystycznych ani modeli uczenia maszynowego - jedynie przekształceń. Łatwo jest to postrzegać jako nie do końca naukę o danych, ponieważ nie wymaga niczego poza dużą ilością arytmetyki, ale często wykonywanie przekształceń we właściwy sposób jest niezwykle cenne. Większość innych osób w firmie nie ma dostępu do danych, nie ma możliwości skutecznego przeprowadzenia transformacji lub nie wie, jakie przekształcenia należy wykonać. W zależności od danych możesz chcieć dorzucić niektóre metody statystyczne, takie jak znajdowanie wartości na różnych poziomach percentyla lub obliczanie odchylenia standardowego.

WIZUALIZUJ DANE LUB UTWÓRZ TABELĘ PODSUMOWUJĄCĄ

Po wykonaniu odpowiednich przekształceń utwórz wizualizacje lub tabele podsumowań, aby lepiej zobaczyć, co dzieje się w danych. Kontynuując wcześniejszy przykład, jeśli masz liczbę klientów każdego miesiąca, możesz utworzyć wykres słupkowy, aby zobaczyć, jak się zmienili. Ten wykres może ułatwić zobaczenie, jakie wzorce znajdują się w danych w sposób, w jaki nie można było po prostu wydrukować ramki danych na ekranie. Faktyczna wizualizacja, którą wybierzesz, zależy w dużym stopniu od dostępnych danych. Możesz użyć wykresu liniowego, wykresu pudełkowego lub wielu innych opcji. Możesz również utworzyć tabele danych podsumowujących zamiast wykresu, w zależności od tego, co próbujesz zrozumieć. Zauważ, że podczas tworzenia wizualizacji możesz zdać sobie sprawę, że musisz zmienić niektóre etapy transformacji. Prawdopodobnie będziesz wiele razy przechodził między krokami. Ponieważ będziesz iterować przez wizualizacje i ciągle przekształcać dane, będziesz musiał zrównoważyć chęć usunięcia tych przeciętnych, aby zachować czystość kodu, z chęcią zapisania wszystkiego na wypadek, gdybyś go ponownie potrzebował. Najlepszą praktyką jest zapisanie jak najwięcej, pod warunkiem, że (1) Twój stary kod nie psuje się po wprowadzeniu dalszych zmian, oraz (2) możesz wyraźnie zaznaczyć, które wyniki są „dobre”. Unikaj przechowywania kodu, który nie działa w twojej analizie lub ogromnych obszarów kodu, które są zakomentowane; takie sytuacje sprawiają, że utrzymanie kodu jest niezwykle trudne. To podejście jest dodatkowo ulepszane przez użycie kontroli wersji, takiej jak git i GitHub; dzięki ciągłemu dokonywaniu zatwierdzeń za każdym razem, gdy dodasz nową zawartość do analizy, będziesz w stanie prowadzić dziennik tego, co zrobiłeś i cofnąć kod, który nagle psuje.

STWÓRZ MODEL WEDŁUG POTRZEB

Jeśli zauważysz w swoich danych wzorce, które sugerują, że modelowanie jest dobrym pomysłem, zrób to! Może warto zastosować model szeregów czasowych do liczby klientów, aby na przykład przewidzieć klientów w przyszłym roku. Tworząc modele, warto przedstawić wyniki i zwizualizować je, aby zrozumieć, jak dokładne lub przydatne są te modele. Możesz tworzyć wykresy, które porównują przewidywane wyniki z rzeczywistymi wartościami lub pokazują metryki, takie jak wyniki dokładności i wartości ważności funkcji. Jeśli tworzysz modele uczenia maszynowego, które mogą być używane poza analizą, na przykład przez potencjalne wprowadzenie do produkcji, upewnij się, że izolujesz kod, który buduje model, od ogólnej pracy analitycznej. Ponieważ w przyszłości będziesz chciał używać tylko modelu, będziesz musiał łatwo wyciągnąć ten kod z kodu, który tworzy ogólne wykresy wizualizacyjne.

POWTARZANIE

Te kroki należy wykonać dla każdego punktu planu analizy. Podczas tych kroków możesz mieć nowy pomysł na to, co analizować lub zdać sobie sprawę, że to, co uważałeś za rozsądne pytanie, nie ma sensu. W tym momencie powinieneś dostosować swoją analizę, planuj i kontynuuj swoją pracę. Jest prawdopodobne, że różne punkty planu analizy są ze sobą powiązane, więc kod, którego użyłeś w

jednym punkcie, zostanie powtórzony w innym. Warto włożyć wysiłek w ustrukturyzowanie planu analizy, tak aby można było wielokrotnie uruchamiać ten sam kod, a aktualizacje jednej części planu natychmiast wdrażać w innych. Twoim celem jest stworzenie zestawu kodu, który możesz utrzymywać; możesz go łatwo modyfikować, nie tracąc czasu na śledzenie złożonego kodu.

Ważne punkty do odkrywania i modelowania

Praca nad eksploracją i modelowaniem danych jest bardzo zależna od problemu, który próbujesz rozwiązać. Techniki matematyczne i statystyczne, których używasz do grupowania danych, różnią się znacznie od tych, które służą do prognozowania lub próby optymalizacji decyzji. Biorąc to pod uwagę, przestrzeganie pewnych ogólnych wytycznych może sprawić, że różnica między analizą OK a dobrą może być.

SKUP SIĘ NA ODPOWIEDZI NA PYTANIE

Jak omówiono wcześniej, niezwykle łatwo jest marnować czas na pracę, która nie wspiera celu. Jeśli analizujesz zamówienia klientów, aby sprawdzić, czy możesz przewidzieć, kiedy klient nigdy nie wróci, możesz uzyskać działający model sieci neuronowej, a następnie spędzić wiele tygodni na dostrajaniu hiperparametrów. Jeśli interesariusz chce tylko odpowiedzi „tak" lub „nie", dotyczącej tego, czy model jest wykonalny, dostrojenie hiperparametrów, aby model był nieco bardziej wydajny, nie pomaga. Tygodnie spędzone na dostrajaniu hiperparametrów można było zamiast tego poświęcić na coś bardziej istotnego. Podczas przeprowadzania analizy ważne jest, aby skupić się na planie analizy i odpowiadać na pytanie zadane przez firmę. Oznacza to ciągłe zadawanie sobie pytania „Czy to ma znaczenie?" To pytanie powinno być czymś, co należy rozważyć za każdym razem, gdy tworzysz spisek lub tabelę. Jeśli ciągle myślisz, że to, co robisz, jest istotne, to świetnie. W znacznie bardziej prawdopodobnym przypadku, w którym od czasu do czasu myślisz „Ta fabuła (lub tabela) nie jest przydatna", być może będziesz musiał dostosować swoją pracę. Najpierw spróbuj przerwać to, co robisz i przyjąć inne podejście do problemu. Jeśli próbujesz pogrupować klientów według ich wydatków, spróbuj zamiast tego wykonać grupowanie. Przyjmując radykalnie inne podejście, masz większe szanse na sukces niż tylko nieznacznie zmieniając to, co robisz. Po drugie, porozmawiaj ze swoim menedżerem lub interesariuszem projektu; możliwe, że dane, których używasz, nie są skuteczne w rozwiązaniu danego problemu. W ciągu tygodni, w których przeprowadzasz analizę, powinieneś stale budować zbiór naprawdę istotnych wyników i (najlepiej) postępować zgodnie z planem analizy.

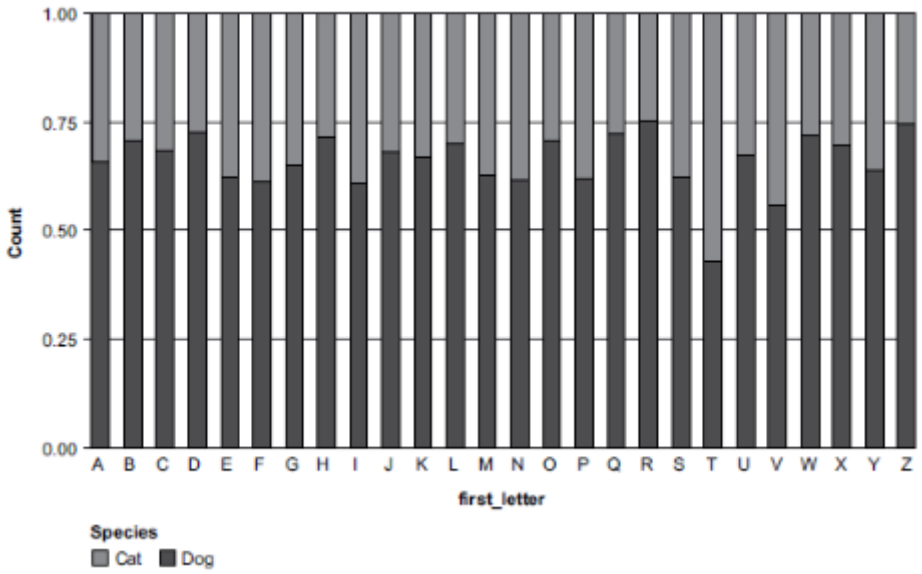
STOSUJ PROSTE METODY ZAMIAST ZŁOŻONYCH

Złożone metody są tak ekscytujące! Po co używać regresji liniowej, skoro można użyć losowego lasu? Po co używać losowego lasu, skoro można korzystać z sieci neuronowej? Wykazano, że metody te działają lepiej niż zwykła regresja lub grupowanie k-średnich i są bardziej interesujące. Więc kiedy ludzie proszą Cię o rozwiązywanie pytań biznesowych za pomocą danych, z pewnością powinieneś dostarczyć najlepsze możliwe metody. Niestety, złożone metody mają wiele wad, których nie widać po skupieniu się wyłącznie na ich dokładności. Kiedy przeprowadzasz analizę, celem nie jest uzyskanie najlepszej możliwej dokładności lub przewidywania; to odpowiedź na pytanie w sposób zrozumiały dla biznesmena. Oznacza to, że musisz wyjaśnić, dlaczego uzyskałeś taki wynik. Dzięki prostej regresji liniowej łatwo jest przedstawić wykresy pokazujące, w jakim stopniu każda cecha przyczyniła się do wyniku, podczas gdy w przypadku innych metod bardzo trudno jest opisać, w jaki sposób model przyniósł wynik, co utrudnia przedsiębiorcom uwierzenie w Twoje wyniki. Bardziej skomplikowane metody są również bardziej czasochłonne w przygotowaniu; dostrojenie i uruchomienie sieci neuronowej zajmuje trochę czasu, podczas gdy regresja liniowa jest dość szybka. Więc kiedy robisz swoją analizę, wybieraj proste metody tak często, jak to możliwe, zarówno w modelach, jak i w transformacjach i agregacjach. Zamiast na przykład przycinać pewien procent wartości odstających,

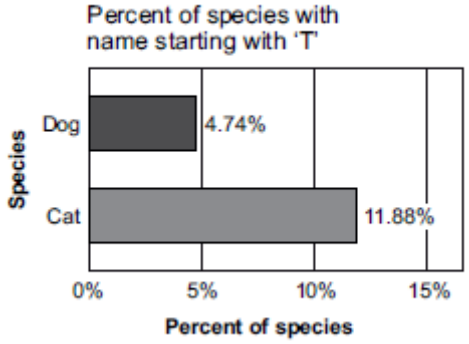
przeprowadź transformację logarytmiczną lub weź medianę zamiast średniej. Jeśli regresja liniowa działa dość dobrze, nie trać czasu na budowanie sieci neuronowej, aby nieznacznie poprawić dokładność. Trzymanie się prostych metod, gdy tylko jest to możliwe, sprawia, że wynik jest znacznie łatwiejszy do zrozumienia dla innych osób, a tobie do obrony i debugowania.

ROZWAŻYĆ WYKRES DO BADAŃ VS WYKRES DO UDOSTĘPNIANIA

Istnieją dwa różne powody, dla których analityk danych wybrałby wizualizację danych: do eksploracji i udostępniania. Kiedy tworzysz wykres do eksploracji, chodzi o to, aby pomóc naukowcom danych zrozumieć, co dzieje się w danych. Posiadanie skomplikowanego i słabo oznaczonego wykresu jest w porządku, o ile badacz danych go rozumie. Kiedy tworzysz wykres do udostępniania, celem jest, aby ktoś, kto nie wie zbyt wiele na temat danych, uzyskał konkretny punkt, który próbuje ustalić analityk danych. Tutaj fabuła musi być prosta i przejrzysta, aby była skuteczna. Dokonując analizy, powinieneś używać wielu wykresów eksploracyjnych, ale tych wykresów nie należy używać do udostępniania. Rozważmy przykład oparty na fikcyjnych danych o imionach zwierząt domowych w mieście: badacz danych chce zrozumieć, czy litera, od której zaczyna się imię zwierzęcia, odnosi się do gatunku zwierzęcia (kot lub pies). Analityk danych ładuje dane i tworzy wizualizację, pokazując dla każdej litery podział kotów i psów, których imiona zaczynają się na tę literę



Jeśli przyjrzysz się uważnie, zauważysz, że słupek T zawiera znacznie większą liczbę kotów niż psów – to znaczące odkrycie dla badacza danych. Biorąc to pod uwagę, ta fabuła nie jest czymś, co chciałbyś pokazać interesariuszowi; w fabule dużo się dzieje i na pierwszy rzut oka nie jest jasne, o co chodzi. Rysunek poniższy pokazuje te same dane wykreślone w inny, bardziej przystępny sposób.



W tej wersji jasne jest, że koty mają 12% szans na otrzymanie imienia zaczynającego się na T, podczas gdy psy mają tylko 5% szansy. Teraz te same dane mogą być udostępniane.

CIĄGLE GOTOWY DO UDOSTĘPNIANIA

Wynik analizy może przybierać różne formy, a wybór formy zależy zazwyczaj od grupy docelowej. Jeśli analiza ma być przekazana przedsiębiorcom, często używany jest pokaz slajdów lub dokument edytowalny. PowerPoint lub Word (lub Prezentacje Google lub Dokumenty Google) to dobry wybór, ponieważ każdy może je wyświetlić (o ile ma pakiet Microsoft Office dla dwóch pierwszych) i może zawierać wiele wykresów, tabel i opisów tekstowych . Jeśli analiza dotyczy osób technicznych, możesz przekazać wyjściowy plik HTML Jupyter Notebook lub R Markdown. Te metody są dobre, ponieważ zwykle wymagają mniej pracy do polerowania (to znaczy, że nie musisz tracić czasu na wyrównywanie figur na slajdzie). Jeśli analiza wymaga przekazania wielu tabel danych finansistom, Excel może być najlepszym wyborem. Excel jest doskonałym narzędziem, gdy użytkownik końcowy musi wziąć liczby w wynikach i wykonać na nich dalsze obliczenia. Powinieneś zdecydować na wczesnym etapie procesu tworzenia analizy, jakiego rodzaju wyniki chcesz dostarczyć, aby uniknąć późniejszych przeróbek. W zależności od tego, jak duży jest zakres analizy, będziesz chciał okresowo kontaktować się z osobą, dla której przeprowadzasz analizę i pokazywać jej swoją pracę. Takie podejście zapobiega okropnej sytuacji, w której spędzasz tygodnie pracując nad analizą w odosobnieniu, a kiedy nadszedł czas, aby ją przekazać, interesariusz wskazuje coś, co unieważnia całą twoją pracę (np. „Przyjrzałeś się sprzedaży klientom, ale zapomniałeś wziąć pod uwagę zwroty."). W takiej sytuacji, gdyby ten element został wskazany na początku, można by uniknąć konieczności wyrzucania dużej ilości pracy. Oprócz unikania złych sytuacji, interesariusz często może wnieść swój wkład, sugerując możliwe obszary, na których należy się skoncentrować, lub metody, które należy wypróbować. W pewnym sensie kontaktowanie się z interesariuszami podczas tworzenia analizy jest podobne do koncepcji zwinnego tworzenia oprogramowania: ciągłe wprowadzanie ulepszeń do pracy, a nie tworzenie jednej ogromnej wersji oprogramowania. Częste kontrole u interesariuszy są świetne, ale naukowcy danych często zaniedbują ich wykonywanie. Wadą odprawy z kimś jest to, że praca musi być pokazana naukowcom, którzy nie zajmują się danymi; musi być na wystarczającym poziomie dopracowania, aby nie krępowało się go pokazywać. Rzeczy takie jak fabuły z wyraźnymi etykietami i znaczeniem, kod z minimalną liczbą błędów i podstawowa historia tego, co się dzieje, są wymagane. Tak więc naukowcom zajmującym się danymi łatwo jest pomyśleć: „Odłożę udostępnianie swojej pracy do czasu, aż ją dopracuję, a doszlifowanie odłożę na później”. Nie rób tego! Na dłuższą metę prawie zawsze będzie to oznaczać więcej pracy. Dzięki ciągłemu utrzymywaniu poziomu dopracowania, dzięki czemu możesz udostępniać swój kod, otrzymujesz lepszy produkt.

URUCHAMIANIE JEDNYM PRZYCISKIEM

Tak jak do załadowania i przygotowania danych powinno wymagać uruchomienia tylko jednego skryptu, tak analiza powinna zostać uruchomiona po naciśnięciu jednego przycisku. W Pythonie oznacza to posiadanie notatnika Jupyter, który automatycznie ładuje dane i wykonuje analizę bez błędów. W języku R miej plik R Markdown, który ładuje dane, analizuje je i wyświetla plik HTML, dokument Word lub prezentację PowerPoint. Przeprowadzając analizę, będziesz chciał uniknąć uruchamiania zbyt dużej ilości kodu poza skryptem lub uruchamiania skryptów niewłaściwie. Te praktyki zwiększają prawdopodobieństwo, że po ponownym uruchomieniu całego skryptu wystąpi błąd. Można zrobić trochę kodowania ad-hoc, ale po prostu upewnij się, że możesz ponownie uruchomić plik bez błędów. Ta praktyka pomoże Ci utrzymać wyniki w ciągłej gotowości do dzielenia się z innymi ludźmi i zapewni, że poświęcisz mniej czasu na poprawianie skryptu pod koniec analizy.

Zawijanie tego

W zależności od interesariusza dla tej analizy, wynik twojego kodu może wystarczyć do zaspokojenia żądania lub być może będziesz musiał pójść dalej i stworzyć ostateczną wersję. Jeśli wymagana jest dopracowana, ostateczna wersja, taka jak prezentacja PowerPoint, może być konieczne dopracowanie końcowego poziomu wykraczającego poza to, co zrobiłeś podczas dokonywania analizy, aby przestrzegać wytycznych dotyczących stylu firmy. Co najważniejsze, musisz stworzyć narrację do końcowego dokumentu, aby osoby, które nie były zaangażowane w pracę, mogły w pełni zrozumieć wnioski z pracy, co zostało zrobione i dlaczego. Stworzenie tej narracji jest pierwszym krokiem w dobrym dokumencie końcowym. Jaką historię zamierzasz opowiedzieć? Jak zamierzasz przedstawić problem, wyjaśnić, w jaki sposób Twoja praca zapewnia rozwiązanie (lub nie) i omówić kolejne kroki? Istnieje wiele sposobów na stworzenie narracji, ale jednym prostym jest zastanowienie się, jak wyjaśnić na głos pracę osobie, która wcześniej jej nie widziała. Pomyśl o historii, którą chciałbyś im opowiedzieć i spróbuj ją opowiedzieć w swoim dokumencie. Wielokrotnie zadawaj sobie następujące pytania: „Czy to, co pokazuję, będzie zrozumiałe dla moich odbiorców?" i „Co mogę zrobić, aby to poprawić?" W końcu dojdziesz do punktu, w którym będziesz zadowolony ze swoich treści. Będziesz także musiał dodać tekst do swojego dokumentu – zwykle w celu wyjaśnienia narracji, którą masz lub dlaczego warto udostępnić każdy wykres. Ponownie postaraj się, aby było to zrozumiałe dla kogoś, kto nie ma kontekstu, który masz. Niech tekst odpowiada na pytanie „W jaki sposób to, co pokazuję, jest przydatne dla firmy?" Różne firmy mają różne standardy dotyczące ilości tekstu, który należy zawrzeć; niektórzy chcą szczegółowych opisów wyjaśniających wszystko, podczas gdy inne firmy są zadowolone z kilku słów. Staraj się popełniać błędy po stronie nadmiernego wyjaśniania, ponieważ później możesz wyciąć treść. Kiedy uważasz, że Twój materiał jest gotowy, będziesz chciał poddać się ocenie wzajemnej, aby sprawdzić, czy nie ma drobnych błędów przed wysłaniem go do interesariuszy. Zastanów się, czy w zespole ktoś, kto zna kontekst pracy, sprawdzi go, aby sprawdzić, czy wszystko ma sens. W zależności od Twojej firmy, Twój przełożony może wymagać, abyś zrobił to z nimi, aby mogli Cię podpisać.

Finalna prezentacja

Po uzyskaniu zgody przełożonego na analizę należy umówić się na spotkanie z interesariuszem, aby osobiście dostarczyć analizę. Podczas tego spotkania będziesz chciał przeprowadzić ich przez każdy komponent, opisując, co zrobiłeś, czego się nauczyłeś i czego zdecydowałeś się nie zajmować. Spędzisz tak dużo czasu z danymi tworząc analizę, że powinieneś czuć się swobodnie wyjaśniając je i odpowiadając na pytania. W zależności od interesariusza możesz znaleźć się w trakcie całej prezentacji, lub osoba może zapisywać pytania do końca. Pytania mogą się wahać od spokojnych i ciekawych („Dlaczego użyłeś zestawu danych X zamiast zestawu danych Y?") po krytyczne i zaniepokojone („Dlaczego te wyniki nie są zgodne z pracą innego zespołu? Czy w Twoim kodzie są błędy?"). Sposób, w jaki powinieneś radzić sobie z pytaniami, pod wieloma względami jest taki sam, jak odpowiadanie

na pytania podczas rozmowy kwalifikacyjnej: bądź na bieżąco z tym, co wiesz, a czego nie. Można powiedzieć, że musisz się temu przyjrzeć. O ile to możliwe, bądź otwarty ze swoim rozumowaniem („Użyliśmy zestawu danych X, ponieważ obejmował on okres, na którym nam zależało"), a gdy czegoś nie wiesz („Nie jestem pewien, dlaczego nie są one zgodne z drugi zespół; zajmę się tym"). Biorąc to pod uwagę, przez większość czasu te spotkania są spokojne i wolne od konfliktów! Bez względu na to, jak dobra jest twoja analiza, nieuchronnie pojawi się pytanie w formie „A co z _____?", w którym puste miejsce jest czymś, na co nie zwróciłeś uwagi w swojej analizie. Ktoś może zapytać: „A co, jeśli w analizie wykorzystasz tylko dane z ostatniego miesiąca?" Jest to naturalne dla ludzi ze względu na naturę nauki o danych: zawsze można znaleźć więcej sposobów na podzielenie danych i pomysłów na to, co może być przydatne. Dzieje się tak szczególnie często w sytuacjach, w których analiza okazała się niejednoznaczna. W takich sytuacjach osoba zgłaszająca prośbę często chce wskoczyć z nadzieją, że coś może nagle okazać się rozstrzygające. Jako naukowiec zajmujący się danymi, najlepszą rzeczą, jaką możesz zrobić w takich sytuacjach, jest próba delikatnego odrzucenia tych żądań. Chociaż od czasu do czasu prośby okazują się przydatne, równie łatwo mogą zakończyć się bez wyciągania nowych wniosków, co powoduje, że tracisz dni czasu na próby ich przepracowania. Jako specjalista ds. danych powinieneś mieć najlepszą wiedzę na temat tego, co miałoby szansę być wartościowe, a jeśli uważasz, że coś nie jest przydatne, możesz wyciągnąć taki wniosek. Często, gdy przeprowadzasz analizę, pytanie biznesowe, które próbujesz rozwiązać, jest tak abstrakcyjne, że nigdy nie możesz udzielić prawdziwie ostatecznej odpowiedzi. I tak jak wtedy, gdy robiłeś analizę i musiałeś unikać próbowania metody po metodzie, aby znaleźć wynik, po analizie musisz wiedzieć, kiedy przestać.

Odłożenie swojej pracy na półkę

Kiedy ostateczna analiza zostanie dostarczona i zatwierdzona, zostaniesz poproszony o szybkie przejście do następnego zestawu pracy, na przykład kolejnej analizy. Jednak zanim to zrobisz, zrobienie kilku małych kroków znacznie ułatwi Ci życie w przyszłości. Istnieje duża szansa, że w pewnym momencie, za kilka miesięcy lub lat, zostaniesz poproszony o ponowne wykonanie analizy z nowszymi danymi. Jeśli poświęcisz trochę czasu na dokumentowanie swojej pracy, wykonanie tej powtórnej analizy będzie znacznie łatwiejsze. Kroki te to :

* Sprawdź dokładnie, czy możesz ponownie przeprowadzić całą analizę. Wcześniej omawialiśmy, jak sprawić, by Twoja analiza była uruchamiana jednym przyciskiem; w tym momencie powinieneś zrobić ostatnią kontrolę, aby sprawdzić, czy analiza nadal działa.

* Skomentuj swój kod. Ponieważ możesz nie patrzeć ponownie na swój kod przez lata, nawet lekkie komentowanie może pomóc Ci zapamiętać, jak używać lub modyfikować kod.

* Dodaj plik README. Plik README to prosty dokument tekstowy opisujący, do czego służy analiza, dlaczego została wykonana i jak ją uruchomić.

* Bezpiecznie przechowuj swój kod. Jeśli korzystasz z git i GitHub, już to zrobiłeś, ale jeśli nie, zastanów się, jak ktoś może uzyskać dostęp do kodu już dawno.

* Upewnij się, że dane są bezpiecznie przechowywane. Sprawdź, czy wszystkie pliki danych są przechowywane w bezpiecznym miejscu innym niż laptop, takim jak usługi w chmurze (na przykład OneDrive, udostępniony dysk sieciowy lub AWS S3). Ponadto zbiory danych przechowywane w bazach danych powinny być najlepiej sprawdzane, aby upewnić się, że nie zostaną usunięte.

* Dane wyjściowe są przechowywane w udostępnionej lokalizacji. Najczęstszym sposobem, w jaki ludzie udostępniają analizy, są załączniki do wiadomości e-mail, ale nie jest to dobry sposób na ich

archiwizację. Umieść swoje wyniki w miejscu, do którego mają dostęp inni członkowie zespołu i osoby z innych części firmy.

Po zakończeniu tej pracy możesz nazwać analizę naprawdę kompletną. W miarę robienia coraz większej liczby analiz, znajdziesz metody i techniki, które najlepiej Ci odpowiadają, i będziesz coraz lepiej i szybciej je wykonywać.