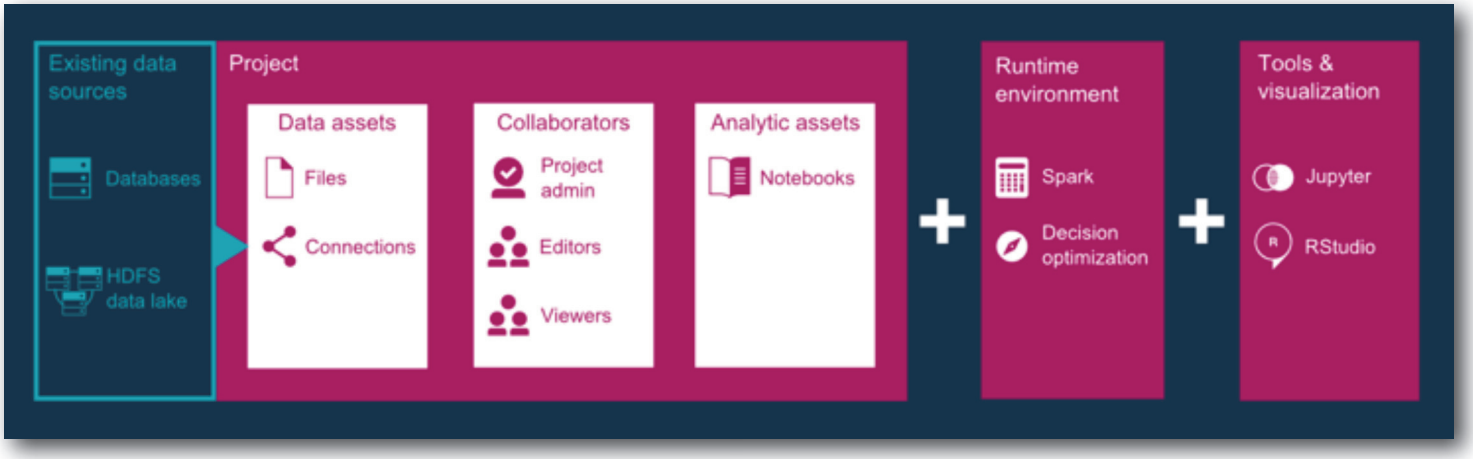
Data Science Experience
wszystkie Twoje narzędzia w jednym miejscu!

Ewa Gruszka
Technical Sales Predictive Analytics (SPSS)

Kim jest Data Scientist?

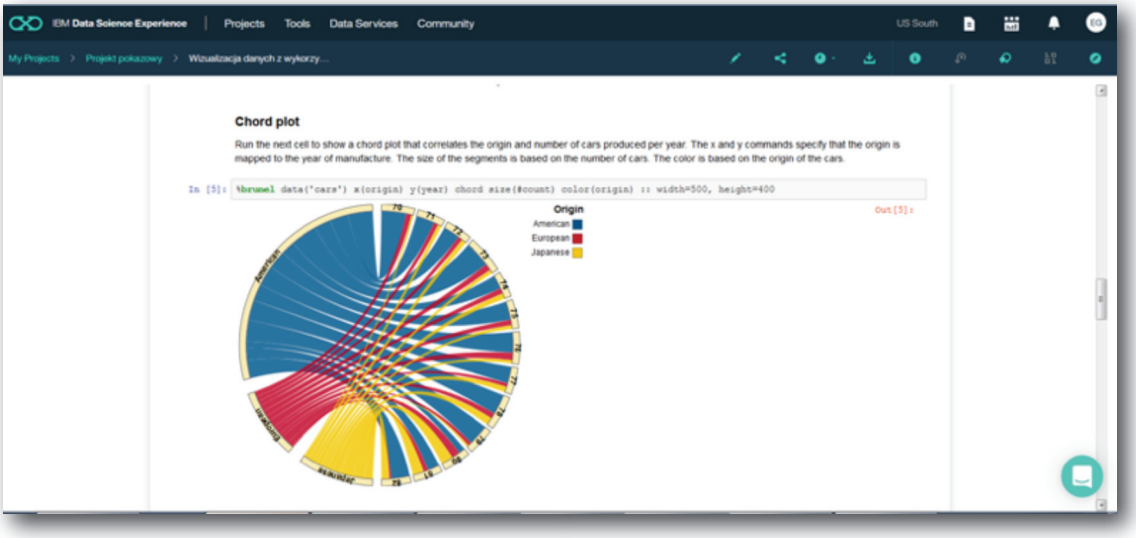
Data Scientist to połączenie kompetencji analityka i informatyka. Jest nim osoba obyta w wielu technologiach, otwarta na zmieniające się trendy i chłonąca najnowsze algorytmy. Taka osoba potrzebuje narzędzi spełniajacych jej wymagania dostępu do zaawansowanej technologii. Z drugiej strony, czy można sobie wyobrazić środowisko zapewniajace wiele języków programowania, z wydajnym silnikiem przetwarzania, z pełną integracją z dowolną bazą danych, udostepniające API do scoringu w trybie batch bądź real time, a do tego dotępne 'z pudełka'? Platforma Data Science Experience łączy w jednym miejscu najbardziej popularne narzędzia open source, dodane unikalne funkcjonalności IBM oraz mechanizmy społecznościowe dzielenia się wiedzą i pracę grupową. Dzięki temu



analitycy otrzymują spójne i zintegrowane środowisko pracy, które mogą wykorzystywać do analizy i rozwiązywania problemów biznesowych. Środowisko to jest uniwersalne i może być wykorzystywane do analizowania dowolnych zagadnień z dowolnego sektora biznesowego. Platforma stanowi jedno spójne środowisko dla użytkowników o różnym poziomie zaawansowania (data scientist, analityk, użytkownik biznesowy), na której możliwa jest płynna współpraca i zarządzanie wspólnymi zasobami. Każdy z nich może korzystać z preferowanych przez niego narzędzi.

Data Scientist - Jupyter jak Julia, Python i R, ale nie zapominajmy o Scala.

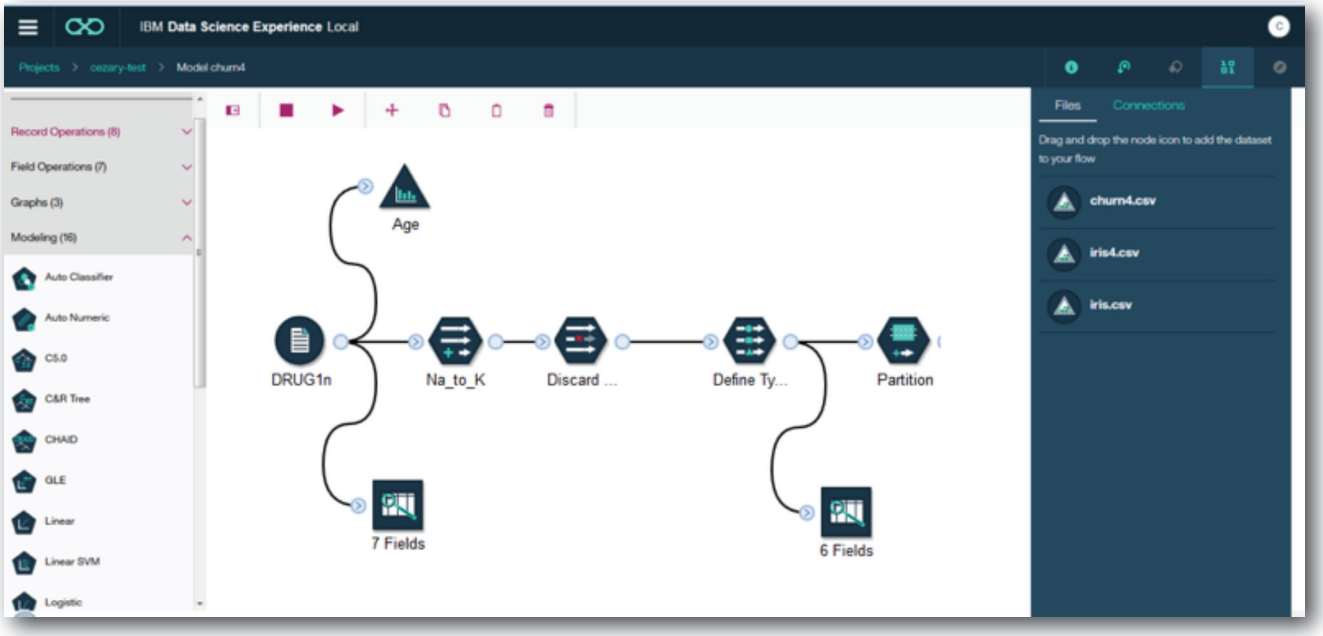
Jupyter Notebooks stworzone do obsługi najbardziej popularnych języków programowania, w IBM Data Science Experience wspierają technologie open source takie jak Python, R oraz Scala. Użytkownicy otrzymują w pełni skonfigurawane środowisko z bardzo dużą liczbą preinstalowanych bibliotek z możliwością rozszerzenia/ doinstalowania innych bibliotek. Na uwagę zasługuje fakt, że dostępne są wszystkie najpopularniejsze biblioteki deep learning t.j.: TensorFlow, Theano, Keras, Lasagne, Caffe. Kod pisany w tym narzędziu jest wykonywany bezpośrednio na klastrze Sparka, a jego wywołanie może nastąpić linijka po linijce, bądź też według zadanego wcześniej harmonogramu.



Jeżeli istnieje potrzeba udostępniania wyników analiz w przystępnej graficzniej formie, to DSX daje możliwość zbudowania ich w oparciu o RStudio i framework Shiny. Dzięku temu, użytkownicy biznesowi mogą podejmować decyzje na podstawie zbudowanych w środowisku interaktywnych aplikacji.

Analityk: budowa modeli predykcyjnych bez znajomości języków programownaia

Canvas to moduł wzorowany na oprogramowaniu IBM SPSS Modeler. Zapewnia graficzne, intuicyjne środowisko wspierające analitykę na każdym etapie pracy. Przyjazny interfejs umożliwia tworzenie analitycznych schematów przetwarzania danych za pomocą powiązanych ze sobą predefiniowanych węzłów, które wspierają różne operacje wykonywane na danych. Dzięki temu użytkownik może budować najbardziej skomplikowane procesy przetwarzania danych bez napisania linijki kodu. Dodatkowo, do wyboru ma szeroki wahlarz gotowych, ale edytowalnych węzłów do budowy modeli predykcyjnych. Co ważne, środowisko integruje się z silnikiem przetwarzania Spark.



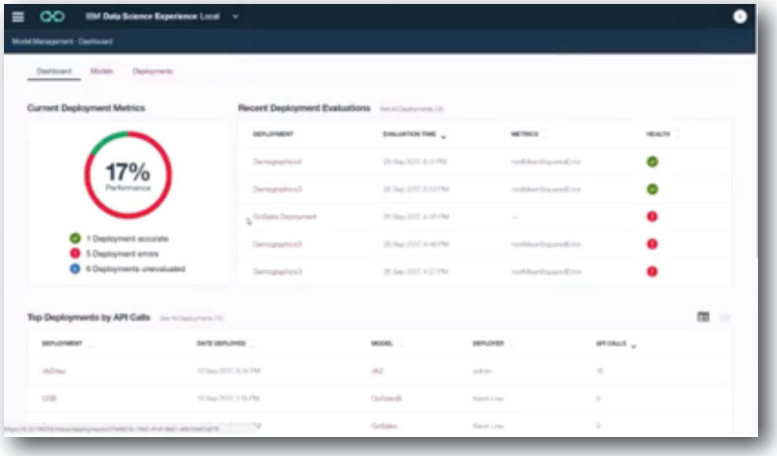
Użytkownik biznesowy: Model predykcyjny na jeden klik

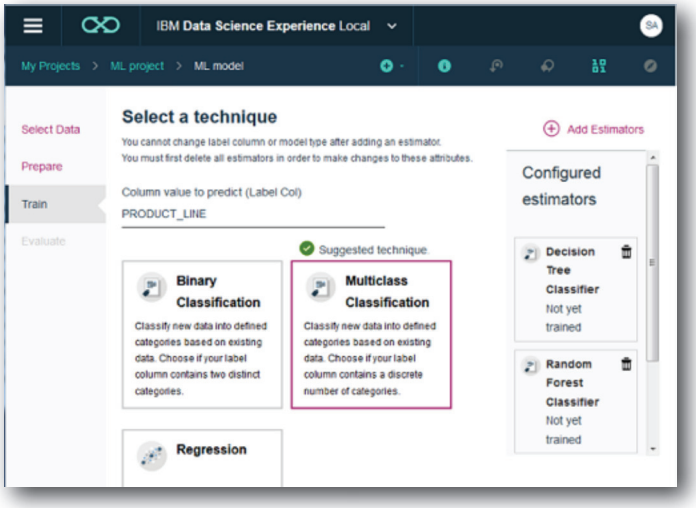
Dla użytkowników nie posiadających szerokiej wiedzy na temat analizy danych i modelowania ‘data mining’ został przygotowany interaktywny asystent. Jest to mechanizm do automatycznego przetwarzania danych, a następnie budowania modeli predykcyjnych w oparciu o algorytmy udostępniane przez IBM Watson Machine Learning. Przygotowanie próbki, wybór transformacji, a następnie technik uczących może odbywać się w pełni automatyczny sposób bądź za pomocą intuicyjnego kreatora. Mechanizm poprowadzi użytkownika ‘za rączkę’ przez proces uczenia, ewaluacji i uprodukcyjniania modelu, w tym przez udostępnianie jego przeliczeń w formie API do wywołań w czasie rzeczywistym.

Jak połączyć pracę różnych użytkowników?

Projekty są podstawowym zasobem do kolaboracji pomiędzy wieloma osobami w organizacji. Ułatwiają współpracę stanowiąc repozytorium danych, zdefiniowanych połączeń oraz obiektów tj. Notebooków, projektów Canvas czy modeli wytworzonych w oparciu o IBM Watson Machine Learning. Oznacza to, że współpracownicy o różnym poziomie zaawansowania nadal mogą wymieniać sie swoimi zasobami wytworzonymi za pomocą różnych narzędzi! Udostępnione wyniki pracy w postaci gotowego modelu predykcyjnego, mogą być uruchamiane w ramach projektu w trybie wsadowym według harmonogramu, lub czasie rzeczywistym za pomocą gotowego API. Dzięki udostępnieniu modeli online, mogą z nich korzystać dowolne zewnętrzne systemy, a wywołanie może nastapić np. z aplikacji bądź z Jupyter Notebooks dostępnych na platformie. Kontrola wdrożeń odbywa się za pomocą wygodnej konsoli do zarządzania zbudowanymi i przeliczanymi modelami. Mechanizm ten obejmuje nie tylko modele przygotowane w DSX, ale także w innych narzędziach i zaimportowanych pod postacią plików: PMML, .gz, .jar itp.

We współpracy pomiędzy użytkownikami ważny apsekt stanowi zarządzanie uprawnieniami. Na poziomie każdego z powyższych obiektów, w zależności od zaawansowania współpracy, mogą zostać nadane różne poziomy dostępu (odczyt, edycja, administrator itp).





Gdzie instalować?

Odpowiedź jest prosta - wszędzie! IBM Data Science Experience jest rozwiązaniem z pudełka, które można zainstalować na dowolnej architekturze: dedykowanym serwerze w firmie, na desktopie bądź może być dostępne bezpośrednio w chmurze. To ostatnie rozwiązanie oznacza pewność najnowszej wersji bez konieczności utrzymywania środowiska.

Jaka jest wartość dodana przez IBM?

IBM Data Science Experience to nie tylko narzędzia open source, ale też ich pełna integracja z narzędziami oferowanymi przez IBM jak Canvas czy Watson Machine Learning. DSX to gotowe środowisko, bez potrzeby konfiguracji czy doinstalowywania poszczególnych komponentów, a wydajność przetwarzania jest osiągnieta przez zastosowanie silnika Spark.

Projekty zapewniają możliwość wymiany zasobów i współpracy przy wykonywaniu zadań. Analityka może odbywać się na gotowych i zaimportowanych zasobach z plików lokalnych bądź wytworzonych bezpośrendio w środowisku narzędziach, o różnym stopniu zaawansowania. DSX to także gotowe konektory do baz on-premise np. DB2, Oracle, SQL Server itp., jak i w chmurze przykładowo Db2 Warehouse on Cloud, Amazon S3, Cloudant, PostgreSQL i wiele innych. Środowisko zapewnia gotowe API oraz szereg preinstalowanych bibliotek zarówno do wizualizacji (Brunel, PixieDust) jak i do przetwarzania (np. PySpark) oraz maszynowego uczenia (np. SparkML), w tym algorytmy deep learningowe, oczywiście z możliwością doinstalowania własnych. Data Science Experience to zaawansowane środowisko do analityki gotowe do użytkowania.