Optymalizacja hiperparametrów polega raczej na wynikach eksperymentów niż na teorii.

William Koehrsen, Data Scientist w Cortex Intel

Rozdział 10

Optymalizacja i wdrożenie modeli

Nasz ostatni wspólny eksperyment będzie okazją do praktycznego zastosowania informacji z poprzednich rozdziałów. Budując model predykcyjny, tym razem skoncentrujemy się na jego optymalizacji, przede wszystkim na dobraniu optymalnych wartości hiperparametrów. Ponieważ będziemy pracować z dużym zbiorem danych, liczącym około miliona przykładów opisanych ponad setką zmiennych, wszystkie operacje wykonamy po stronie serwera SQL. Pozwoli to utrwalić znane i poznać nowe metody integracji serwera SQL z językami R i Python. Na końcu udostępnimy użytkownikom gotowy model i wyniki jego predykcji.

Jak zwykle, eksperyment przeprowadzimy zgodnie z wytycznymi metodyki CRISP-DM. Jednak w tym przypadku Państwa zadanie nie ograniczy się do zbudowania modelu według przedstawionych wskazówek. Chociaż byłby to działający model, to odpowiednio stosując poznane techniki uczenia maszynowego, będą Państwo mogli znacząco poprawić jego jakość. Zaczynajmy.

10.1. Zrozumienie problemu

Gdybyśmy mieli kryształową kulę, pożyczalibyśmy pieniądze tylko tym, którzy nam je zwrócą. Firmy finansowe mogą skorzystać z analizy predykcyjnej, aby ograniczyć liczbę pożyczek oferowanych kredytobiorcom, którzy najprawdopodobniej ich nie spłacą. Ponieważ zwiększa to ich zyskowność, dzisiaj prawie wszystkie firmy finansowe przeprowadzają taką ocenę ryzyka kredytowego.

Naszym zadaniem jest zbudowanie modelu oceniającego ryzyko kredytowe dla fikcyjnej firmy pożyczkowej ze Stanów Zjednoczonych. Dysponuje ona historycznymi danymi o udzielonych kredytobiorcom pożyczkach obejmującymi informację o spłatach i statusie poszczególnych kredytów (rys. 10.1).

Rysunek 10.1. Najwięcej pożyczek jest na bieżąco spłacanych (status *Current*) lub zostało spłaconych w terminie (status *Fully Paid*). Pozostałe statusy świadczą o problemach ze zwrotem kredytów, co oznacza, że firma ma problem z odzyskaniem ponad 800 mln USD pożyczonych klientom

Naszym zadaniem jest zbudowanie modelu, który oceni kredyt jako dobry lub zły. Taki model będzie używany nie tylko do oceny udzielanych kredytów, ale również do analiz typu *Co by było, gdyby?*. Firma planuje wykorzystać nasz model do sprawdzania, jak zwiększenie oprocentowania pożyczek wpłynęłoby w różnych stanach na ich spłacalność. Budując model, powinniśmy pamiętać, że na serwerze SQL, który będzie używany do predykcji, nie będzie zainstalowanej usługi SQL Server Machine Learning.

Mamy więc zbudować model klasyfikatora, który będzie dzielił pożyczki na dobre (spłacane w terminie) i złe (pozostałe). Finalny model wdrożymy na serwerze SQL 2019 bez zainstalowanej usługi uczenia maszynowego.

10.2. Zrozumienie i przygotowanie danych

Przykładowe dane pochodzą ze strony `https://www.lendingclub.com/info/download-data.action`. Są to dane o pożyczkach udzielonych w latach 2017–2018. Proszę zwrócić uwagę,