Analiza Danych

Projekt „Pożyczki” - Raport

Kamil ciosek

Oskar Kanigowski

Nina Górczyńska

Styczeń 2025

Opis Bazy Danych

Pożyczki

Spośród wszystkich branż, domena ubezpieczeniowa ma największe zastosowanie analityki i metod nauki o danych. Ten zestaw danych zapewni ci wystarczający przedsmak pracy nad zestawami danych z firm ubezpieczeniowych, z jakimi wyzwaniami się mierzą, jakie strategie są stosowane, jakie zmienne wpływają na wynik itp. Jest to problem klasyfikacji. Dane mają 615 wierszy i 13 kolumn.

Problem

Firma chce zautomatyzować proces kwalifikowania się do pożyczki (w czasie rzeczywistym) w oparciu o dane klienta podane podczas wypełniania wniosku online. Dane te obejmują płeć, stan cywilny, wykształcenie, liczbę osób na utrzymaniu, dochód, kwotę pożyczki, historię kredytową i inne. Aby zautomatyzować ten proces, firma postawiła sobie za cel zidentyfikowanie segmentów klientów, którzy kwalifikują się do otrzymania kwoty pożyczki, aby móc skierować do nich konkretną ofertę. W tym przypadku dostarczono częściowy zestaw danych.

Opis danych

Baza składa się z 13 kolumn, a te z kolei zawierają następujące wartości:

1. Loan\_ID - indywidualny numer ID każdej obserwacji
2. Gender - informuje jaka płeć jest przyjęta w danej obserwacji (male/female)
3. Married - dane, czy osoba jest w związku małżeńskim (yes/no)
4. Dependents - Osób zależnych (wartości liczbowe, 0,1,2 lub 3+)
5. Education - informacja, czy dana osoba ukończyła studia (graduate, not graduate)
6. Self\_Employed - informuje, czy osoba jest samozatrudniona (Yes/No)
7. ApplicantIncome - dochód wnioskodawcy o pożyczkę (wartości liczbowe)
8. CoapplicantIncome - dochód współwnioskodawcy o pożyczkę (wartości liczbowe)
9. LoanAmount - kwota pożyczki do wzięcia (?) (wartości liczbowe)
10. Loan\_Amount\_Term - czas na jaki wzięta będzie pożyczka (wartości liczbowe)
11. Credit\_History - historia kredytowa (czy ktoś brał wcześniej kredyt? Wartości 0/1)
12. Property\_Area - obszar nieruchomosci (?)
13. Loan\_Status - status kredytu (Y/N) - czy uzyskany czy nie
14. Data Wrangling

Na początku pracy z zadaną bazą danych, musimy zrobić wstępną analizę oraz przekształcić dane tak, by móc wydobyć z nich jak najwięcej informacji.  
Poniżej przedstawiamy proces naszej obróbki danych.

1. Analiza danych

Na tym etapie sprawdzamy jakie mamy zmienne oraz jaki charakter one przyjmują. Sprawdzamy, czy w naszej Bazie nie ma braków danych oraz czy dane są jednolite (Yes/No, 0/1 itd.). Oprócz tego oceniamy, które zmienne są przydatne, a z jakich możemy zrezygnować.

Poniżej przedstawiamy proces **Wstępnej Analizy Danych** wraz z podsumowaniem odnośnie napotkanych problemów.

2. Czyszczenie oraz standaryzacja danych

Po wstępnej analizie dokonujemy czyszczenia oraz standaryzacji naszych danych. Usuwamy komunę „Loan\_ID”, która nie wnosi istotnych informacji do analizy końcowej.

Standaryzacja danych polega na przekształceniu zmiennych z tekstowych na liczbowe (np. Married; Yes = 1, No =0). Takie typy zmiennych są przydatne do np. analizy statystycznej.

3. Imputacja danych

Po ustandaryzowaniu danych, przyszedł czas na uzupełnienie braków.

Dla braków w zmiennej „married” stosujemy zastąpienie braków danych medianą wyliczoną z danych zawartych w bazie.

Przy pozostałych danych zdecydowaliśmy się na metodę „mice”. Ta metoda tworzy możliwe uzupełnienia dla brakujących danych, opierając się na wzorcach w kompletnej bazie.

4. Wizualizacja przygotowanych danych

Na tym etapie tworzymy ramkę danych, w której znajdą się nasze dane, które po standaryzacji oraz imputacji otrzymują odpowiednie etykiety w celu zachowania spójności w analizie. Tzn., że dane dalej będą przyjmować wartości numeryczne, ale zmiana etykiet pomoże w prezentacji analizowanych danych (gender zmienione na 0-1, a etykieta pozostaje jako „male” oraz „female”.).