Inżynieria Uczenia Maszynowego

Etap I

Kacper Maj

Łukasz Topolski

1.Opis problemu

Nasi konsultanci narzekają, ponieważ nie są w stanie rozwiązać problemów logistycznych związanych z przyjmowaniem zwrotów zakupionych u nas towarów. Potrzebny jest system, który będzie w stanie powiedzieć jakie zakupione przez klientów produkty zostaną zwrócone.

2. Definicja zadania

Biznesowego:

Ocena czy zakupiony produkt zostanie zwrócony.

Analitycznego:

Przygotowanie modelu klasyfikacji binarnej na podstawie dostarczonych danych.

3. Działanie modelu

Model będzie przewidywał prawdopodobieństwo zwrotu użytkownika na podstawie:

- miejscowości – osoby z miejscowości z gorszą logistyką poczty prawdopodobnie będą dokonywać bardziej przemyślane zakupy, aby uniknąć chodzenia na pocztę lub do paczkomatu
- kategorii produktu – pewne kategorie mają większe szanse na wadliwe egzemplarze lub są bardziej narażone na zepsucie
- model produktu – produkt może mieć wadę projektową
- cena – tańsze produkty tej samej kategorii mogą mieć częstsze wadliwe egzemplarze
- historia sesji użytkownika – na podstawie historii użytkownika powinniśmy być w stanie określić jak często dany użytkownik dokonuje zwrotów również, gdy np. zwykły użytkownik kupuje dwa monitory, to jest większa szansa, że jeden z nich zwróci
- zniżka – z pewnością wpływa ona na zakup, lecz może także wpływać na zwrot, ponieważ produktu mogły być kupione pod wpływem impulsu

Model musi być otwarty na modernizację na nowych danych wygenerowanych podczas operowania modelu; model powinien być otwarty na dodanie nowych kategorii produktów nawet takich, które nie są możliwe do zwrotu po odpakowaniu.

4. Założenia

- Sklep posiada tylko jeden magazyn
- Nasz model pokazuje tylko predykcję czy produkt zostanie zwrócony, opłacalność sprzedaży nie jest istotna
- Sklep nie oferuje specjalnych zwrotów, zwrotów można dokonać tylko 14 dni od otrzymania produktu lub na rękojmi

5. Kryteria sukcesu

Biznesowe:

Załóżmy, że w ciągu miesiąca na 1000 zamówień 100 zostaje zwróconych. Model powinien wykryć prawidłowo co najmniej 85 zwrotów i spośród 900 zamówień niezwróconych przypisać prawidłowo co najmniej 815 zamówień.

Analityczne:

Tworząc nasz model chcielibyśmy uzyskać wyniki na co najmniej następujących poziomach:

- Accuracy: 90%
- Precision: 50%
- Recall: 85%
- F1 score: 63%

6. Ocena danych

Pierwotnie dostarczone dane mogłyby powodować problemy z uczeniem modelu ze względu na następujące czynniki:

- Braki w atrybutach
- Przypadki niemożliwe do wystąpienia (np. ujemna cena produktu)
- Niereprezentatywność zbioru (np. brak danych z listopada i grudnia)

Obecny zbiór danych ma następujące cechy:

- dane są prawdopodobne -nie występują ujemne ceny, itp. (poza jedną nazwą produktu, ale ten błąd nie jest krytyczny)
- nie występują tu braki danych
- na zwrot zamówienia wpływ prawdopodobnie mają następujące atrybuty: miesiąc składania zamówienia, miasto, kategoria produktu, nazwa produktu, firma kurierska, czas dostawy
- potencjalny problem przy uczeniu na podstawie tych danych może stanowić: znacząca dominacja jednej kategorii produktów nad innymi, mała różnorodność miejsc zamieszkania użytkowników, mała liczba zamówień z grudnia