Predykcja rezygnacji klientów z subskrybcji w serwisie telekomunikacyjnym

Opis problemu

W dzisiejszych czasach branża telekomunikacyjna stoi przed wyzwaniem utrzymania lojalności klientów w obliczu rosnącej konkurencji i coraz wyższych oczekiwań konsumentów. Firmy telekomunikacyjne oferujące usługi subskrypcyjne muszą stawić czoła problemowi churnu, czyli odejść klientów. Zjawisko to polega na rezygnacji klientów z dalszego korzystania z usług danego operatora, co bezpośrednio przekłada się na straty finansowe oraz spadek przychodów. Churn może być wynikiem różnych czynników, takich jak niska jakość obsługi, niewystarczająca oferta, problemy techniczne, czy atrakcyjniejsze oferty konkurencji. W związku z tym, skuteczna predykcja odejść klientów staje się kluczowym elementem strategii zarządzania relacjami z klientem (CRM).

Model predykcji odejść klientów może być niezwykle przydatny dla menedżerów ds. marketingu, analityków danych oraz zespołów obsługi klienta w firmach telekomunikacyjnych. Menedżerowie ds. marketingu mogą wykorzystać wyniki modelu do lepszego targetowania kampanii marketingowych oraz tworzenia bardziej atrakcyjnych ofert dla klientów zagrożonych odejściem. Analitycy danych mogą użyć modelu do identyfikacji kluczowych czynników wpływających na churn oraz do przewidywania przyszłych trendów. Zespoły obsługi klienta mogą na podstawie predykcji proaktywnie kontaktować się z klientami, oferując im rozwiązania problemów lub specjalne rabaty, co może przyczynić się do zwiększenia satysfakcji klientów i zmniejszenia liczby odejść.

Problem predykcji odejść klientów jest interesujący z kilku powodów. Po pierwsze, churn ma bezpośredni wpływ na wyniki finansowe firm telekomunikacyjnych, dlatego zrozumienie tego zjawiska i możliwość jego przewidywania ma ogromne znaczenie biznesowe. Po drugie, problem ten wymaga zaawansowanych technik analizy danych i modelowania statystycznego, co stanowi wyzwanie dla analityków danych i może prowadzić do rozwoju nowych metod i narzędzi analitycznych. Po trzecie, skuteczne zarządzanie churnem może znacząco poprawić relacje z klientami oraz ich lojalność, co jest kluczowe w długoterminowej strategii rozwoju firmy. Ostatecznie, praca nad tym

problemem może przynieść wymierne korzyści zarówno dla firm telekomunikacyjnych, jak i dla ich klientów, poprzez lepsze dopasowanie usług do potrzeb konsumentów.

Dane

Dane pochodzą z otwartego zasobu udostępnionego przez University of California, Irvine, znanego ze swojej renomowanej kolekcji zestawów danych do celów badawczych i edukacyjnych. Dataset, pod nazwą "Iranian Churn Dataset," obejmuje dane zebrane przez irańską firmę telekomunikacyjną w okresie dwunastu miesięcy. Zawiera 13 cech i 3 150 obserwacji, co zapewnia wystarczającą próbę do przeprowadzenia analizy. Dane są udostępnione na licencji Creative Commons Attribution 4.0 International (CC BY 4.0), co pozwala na ich swobodne wykorzystanie, pod warunkiem odpowiedniego przypisania źródła. Brak brakujących wartości w zbiorze danych zwiększa jego wiarygodność i użyteczność do celów analitycznych. Głównym problemem zbioru danych jest jego silne niezbalansowanie. Liczba klientów, którzy nie zrezygnowali z subskrypcji (klasa 0), jest 4.25 razy większa od liczby klientów, którzy zrezygnowali z subskrypcji (klasa 1).

Zbiór danych zawiera następujące cechy:

1. **Call Failure (Nieudane Połączenia)** - liczba nieudanych prób połączeń (wartości całkowite)
2. **Frequency of SMS (Częstotliwość wysyłania SMS-ów)** - liczba wysłanych SMS-ów (wartości całkowite)
3. **Complaints (Czy Złożono Skargi)** - informacja, czy klient złożył skargę (wartości binarne)
4. **Distinct Called Numbers (Liczba Odrębnych Połączeń)** - liczba unikalnych numerów, z którymi klient się kontaktował (wartości całkowite)
5. **Subscription Length (Długość Abonamentu)** - czas trwania subskrypcji (wartości całkowite)
6. **Age (Wiek)** - wiek klienta (wartości całkowite)
7. **Age Group (Grupa Wiekowa)** - kategoria wiekowa (wartości całkowite)
8. **Charge Amount (Wysokość Opłaty)** - suma opłat poniesionych przez klienta (wartości całkowite)
9. **Tariff Plan (Rodzaj Usługi)** - rodzaj taryfy, z której korzysta klient (wartości całkowite)
10. **Seconds of Use (Sekundy Użytkowania)** - łączny czas rozmów w sekundach (wartości całkowite)
11. **Status** - aktualny status klienta (wartości binarne)

12. **Frequency of Use (Częstotliwość Użytkowania)** - częstotliwość korzystania z usług (wartości całkowite)
13. **Customer Value (Wartość dla Klienta)** - wartość wyrażona w jednostkach ciągłych

Etykieta:

- **Churn (Czy Zrezygnował z Subskrybcji)** - informacja, czy klient zrezygnował z subskrypcji (wartości binarne)

Dane zawierają informacje zagregowane z pierwszych dziewięciu miesięcy oraz etykiety rezygnacji, które odzwierciedlają stan klientów na koniec dwunastego miesiąca.

Dane te są kluczowe dla stworzenia modelu predykcji odejść klientów, ponieważ zawierają istotne informacje o zachowaniach klientów oraz ich interakcjach z usługami telekomunikacyjnymi. Analizując takie cechy jak liczba nieudanych połączeń, częstotliwość wysyłania SMS-ów, liczba unikalnych numerów, czas trwania subskrypcji, czy liczba złożonych skarg, można zidentyfikować wzorce i czynniki, które wpływają na decyzję klientów o rezygnacji z usług.

Na przykład, częsta zmiana taryf lub wysoka liczba skarg mogą wskazywać na niezadowolenie klienta, które może prowadzić do odejścia. Z kolei dane dotyczące długości subskrypcji i wartości klienta mogą pomóc w zrozumieniu, jakie grupy klientów są bardziej narażone na churn.

Korzystając z tych danych, można opracować model predykcyjny, który pozwoli na wczesne wykrycie klientów zagrożonych odejściem, co umożliwi firmie podjęcie działań zapobiegawczych, takich jak oferowanie specjalnych promocji, poprawa jakości obsługi, czy personalizacja ofert. W rezultacie, efektywne zarządzanie churnem może prowadzić do zwiększenia lojalności klientów i optymalizacji przychodów firmy.

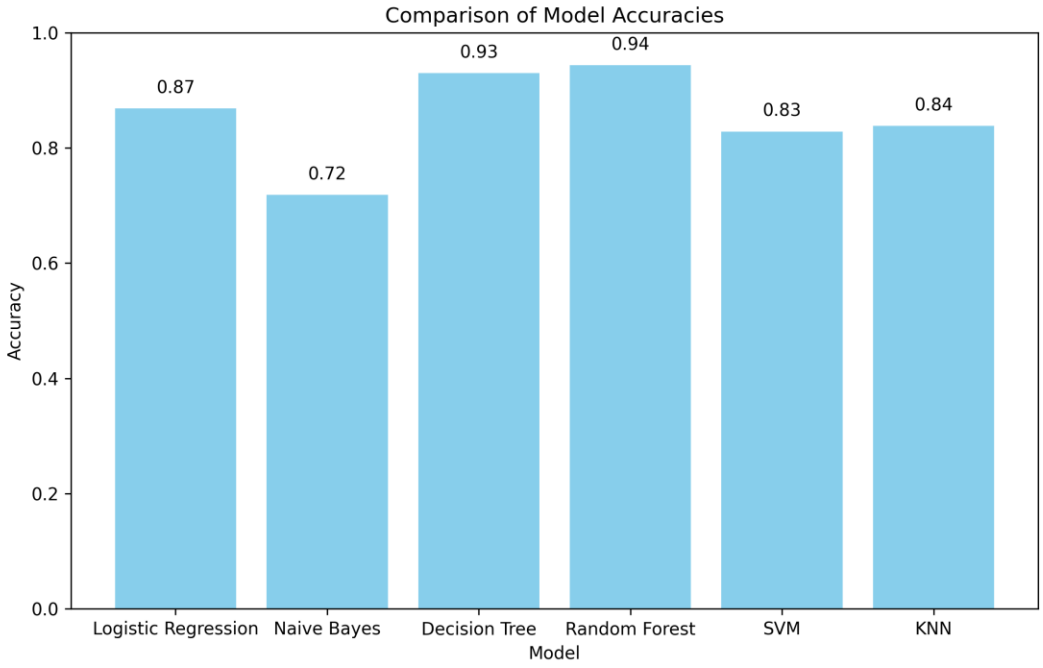
Sposób rozwiązania problemu

Do rozwiązania problemu wybraliśmy model LightGBM (Light Gradient Boosting Machine). Jest to model uczenia maszynowego z kategorii **ensemble learning**, **boosting algorithms**. Stpsując technikę zespołową łączy on wiele słabych modeli drzewiastych w celu stworzenia silnego modelu.

LightGBM (LGBM) jest dobrym wyborem do modelowania w przypadku posiadania niezbalansowanego zbioru danych z kilku kluczowych powodów:

14. **Waga klas**: LightGBM umożliwia przypisanie różnej wagi klasom, co pomaga modelowi lepiej radzić sobie z niezbalansowanymi danymi. Dzięki temu można zwiększyć wagę rzadszej klasy (np. klientów, którzy rezygnują z subskrypcji) i zrównoważyć wpływ obu klas na model.
15. **Boosting drzew**: Jako algorytm oparty na boosting drzew, LightGBM iteracyjnie poprawia swoje predykcje, skupiając się na przykładach, które były wcześniej błędnie klasyfikowane. W przypadku niezbalansowanych danych algorytm ten może skupić więcej uwagi na trudniejszych, mniej reprezentowanych klasach, co prowadzi do poprawy ogólnej skuteczności modelu.
16. **Efektywność obliczeniowa**: LightGBM jest znany ze swojej szybkości i efektywności, zarówno pod względem czasu treningu, jak i zużycia pamięci. Jest to istotne przy pracy z dużymi zbiorami danych, gdzie szybkość iteracji może być kluczowa dla eksperymentowania z różnymi konfiguracjami modelu i przetwarzania dużych wolumenów danych.
17. **Zaawansowane techniki obcinania**: LightGBM stosuje techniki takie jak "Gradient-based One-Side Sampling" (GOSS) i "Exclusive Feature Bundling" (EFB), które pomagają zwiększyć efektywność modelu przy jednoczesnym utrzymaniu wysokiej jakości predykcji. Dzięki temu model jest bardziej precyzyjny nawet w przypadku niezbalansowanych danych.
18. **Parametry regularyzacji**: LightGBM oferuje szeroki zakres parametrów regularyzacji, które można dostosować, aby poprawić wydajność modelu na niezbalansowanych danych, zapobiegając jednocześnie przetrenowaniu.

Model ten wybraliśmy po wcześniejszym wypróbowaniu kilku modeli i wybraniu tego, który osiągnął najbardziej zadowalające nas rezultaty.



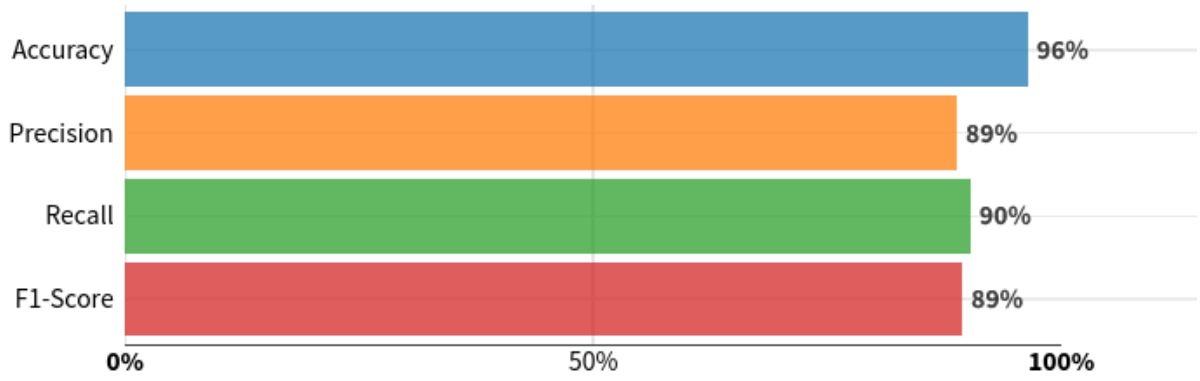
Ze względu na niezbalansowanie klas naszą główną miarą ewaluacji modelu była macierz

pomyłek.

Confusion matrix

Display: % of actual classes

	Predicted 1	Predicted 0	Total
Actually 1	97 %	3 %	100 %
Actually 0	3 %	97 %	100 %



Etapy realizacji projektu

1. Czyszczenie danych:

- Sprawdzenie konsystencji typów danych w kolumnach.
- Upewnienie się, że nie ma brakujących wartości.
- Znalezienie duplikatów.
- Znalezienie wartości odstających (outlierów) przy użyciu metod > 1,5 IQR i > 3 IQR.
- Sprawdzenie konsystencji danych w kolumnach pod kątem zakresu wartości.

Usuwanie wartości odstających i duplikatów.

2. Eksploracja danych:

- Analiza danych w celu zrozumienia ich charakterystyki, rozkładu, zależności między zmiennymi oraz identyfikacji potencjalnych problemów.
- Wyciągnięte wnioski są kluczowe dla dalszego wyboru i customizacji modelu.

2a. Dysbalans klas:

- Sprawdzenie, czy występuje niezbalansowanie klas.

2b. Podstawowa analiza statystyczna:

- Wygenerowanie i opisanie statystyk dla każdej z cech: średnia, mediana, odchylenie standardowe, kwartyle, minimalne i maksymalne wartości.
- Wygenerowanie i wizualizacja rozkładów zmiennych: histogramy i wykresy pudełkowe.

2c. Dalsza wizualizacja zależności między zmiennymi:

- Zobrazowanie zależności między parami zmiennych (scatter plots).
- Mapy ciepła korelacji.
- Identyfikacja rodzaju zależności (liniowe czy nieliniowe).
- Analiza wpływu zmiennych niezależnych na zmienne zależne.

2d. Analiza zmiennych kategorycznych:

- Agregacja danych względem zmiennych kategorycznych, aby zrozumieć różnice między grupami.

2e. Tendencje czasowe (sezonowość):

- Analiza, czy dane wykazują określoną tendencję czasową lub sezonowość.

3. Przygotowanie danych do modelowania: 3a. Normalizacja lub standaryzacja danych:

- Rozważenie potrzeby normalizacji lub standaryzacji danych w zależności od wymagań modelu.

3b. Radzenie sobie z dysbalansem klas:

- Zaimplementowanie metod radzenia sobie z dysbalansem klas, takich jak oversampling i undersampling.

4. Wypróbowanie prostych modeli:

- W celu lepszego zrozumienia danych przetestowanie prostych modeli, takich jak:
    - Regresja logistyczna
    - Drzewo decyzyjne
    - K-Nearest Neighbors (KNN)
    - Naiwny Bayes
    - Maszyny Wektorów Nośnych (SVM)

5. Trening właściwych modeli i dostosowywanie hiperparametrów:

- Trening modeli Random Forest i Light Gradient Boosting Machine i dostosowywanie hiperparametrów w celu zminimalizowania ilości pomyłek, szczególnie tych dotyczących klasy mniejszościowej

6. Udowodnienie działania modelu:

- Stworzenie zbioru walidacyjnego i wytrenowanie danych na podzbiorze w celu dodatkowego sprawdzenia wyników modelu.
- Wytrenowanie i ewaluacja modelu na pomniejszonym zbiorze danych stosując undersampling, czyli usunięcie losowych próbek z klasy większościowej tak, żebby ilość pozostałych była równa ilości próbek z klasy mniejszościowej.

Dyskusja wyników i ewaluacja modelu

Wyniki modelowania

Model LightGBM został wytrenowany na zestawie danych zawierającym 2678 rekordów, co zajęło jedną minutę i 13 sekund. Model wykorzystuje technikę gradient boosting (gbdt) i składa się z 30 estymatorów oraz 31 liści, przy współczynniku uczenia równym 0.2.

Najważniejsze cechy użyte w modelu to:

- **Status**
- **Complains (Skargi)**
- **Frequency of Use (Częstotliwość Użytkowania)**
- **Call Failure (Nieudane Połączenia)**
- **Seconds of Use (Sekundy Użytkowania)**
- **Customer Value (Wartość Klienta)**

Ewaluacja modelu

Na podstawie macierzy pomyłek możemy ocenić skuteczność modelu w klasyfikacji odejść klientów (churn). Model osiągnął następujące wskaźniki:

- **Accuracy (Dokładność):** 96%
- **Precision (Precyzja):** 89%
- **Recall (Czułość):** 90%
- **F1-Score:** 89%

Macierz pomyłek pokazuje, że model prawidłowo sklasyfikował 97% przypadków, w których klienci rzeczywiście zrezygnowali z subskrypcji (klasa 1), oraz 97% przypadków, w których klienci nie zrezygnowali (klasa 0). Błędne klasyfikacje wynoszą 3% dla obu klas.

Podsumowanie

Co się udało?

Model LightGBM skutecznie przewiduje odejścia klientów w firmie telekomunikacyjnej, osiągając wysoką dokładność (96%) oraz znaczące wskaźniki precyzji (89%) i czułości (90%). Kluczowe cechy takie jak status klienta, liczba skarg oraz częstotliwość użytkowania okazały się istotne dla modelu, co potwierdza trafność wybranych zmiennych. Efektywne wykorzystanie technik radzenia sobie z niezbalansowaniem danych pozwoliło na zrównoważenie wpływu obu klas na wyniki modelu.

Jakie były problemy? Jak je rozwiązaliśmy?

Głównym problemem był silny niezbalansowanie zbioru danych, gdzie liczba klientów, którzy nie zrezygnowali z subskrypcji, była 4.25 razy większa od liczby klientów, którzy zrezygnowali. Problem ten rozwiązaliśmy poprzez zastosowanie technik oversamplingu i undersamplingu, które zrównoważyły liczebność klas. Dodatkowo, model LightGBM pozwolił na przypisanie różnej wagi klasom, co poprawiło skuteczność predykcji.

W jaki sposób może być to wykorzystane/rozwinięte w przyszłości?

Model może być używany przez menedżerów ds. marketingu, analityków danych oraz zespoły obsługi klienta do proaktywnego zarządzania relacjami z klientami. Przyszłe prace mogą obejmować:

- Dalszą optymalizację hiperparametrów modelu w celu zwiększenia jego precyzji.
- Integrację modelu z systemami zarządzania relacjami z klientem, aby automatycznie identyfikować klientów zagrożonych odejściem i podejmować działania zapobiegawcze.
- Analizę dodatkowych cech lub zewnętrznych źródeł danych, które mogą wpłynąć na dokładność predykcji.
- Implementację modeli predykcji na większych zbiorach danych lub w różnych segmentach rynku telekomunikacyjnego w celu zbadania ich ogólnej skuteczności.