22 kwietnia 2024

1 Organizacja kodu

Problem: Brak wyraźnej struktury kodu, mieszanie załadunku danych, przetwarzania, tworzenia modeli i ewaluacji w jednym miejscu.
Sugestia: Organizacja kodu w moduły lub wyraźne sekcje, z osobnymi plikami dla przygotowania danych, definicji modelu, treningu i oceny.

2 Obsługa danych

Problem: Używanie względnych ścieżek przy ładowaniu danych.
Sugestia: Implementacja zarządzania konfiguracją przy użyciu plików `.env` lub plików konfiguracyjnych JSON/XML, gdzie ścieżki i źródła danych są zdefiniowane. Poza tym, kod nie jest podzielony na kilka plików (tak samo jak dane) i wszystko jest trzymane w jednym miejscu

3 Analiza Eksploracyjna Danych (EDA)

Problem: Niewystarczające detale na temat przeprowadzonej analizy eksploracyjnej danych.
Sugestia: Przeprowadzenie dokładnej EDA, w tym statystyczne podsumowania każdej kolumny, obsługa brakujących wartości (lub wspomnienie, że one nie występują), wykrywanie i traktowanie wartości odstających oraz wizualizacja rozkładów danych i relacji między cechami.

4 Inżynieria cech

Problem: Brak wyraźnej wzmianki lub dokumentacji procesu inżynierii cech.
Sugestia: Dokumentacja tworzenia nowych cech, ich uzasadnienie i ocena wpływu na wydajność modelu.

5 Selekcja i walidacja modelu

Problem: Wybór modelu wydaje się arbitralny bez uzasadnienia.
Sugestia: Uzasadnienie wyboru modeli na podstawie charakterystyki problemu i danych. Implementacja solidnej strategii walidacji, np. walidacji krzyżowej, aby niezawodnie ocenić wydajność modelu na różnych podzbiorach danych.

6 Dostrojenie hiperparametrów

Problem: Ograniczone metody dostrojenia hiperparametrów.
Sugestia: Stosowanie bardziej kompleksowych metod optymalizacji hiperparametrów, takich jak przeszukiwanie siatkowe, przeszukiwanie losowe czy zaawansowane techniki takie jak optymalizacja bayesowska.

7 Metryki oceny

Problem: Podstawowe implementacje metryk oceny bez nakreślenia, która jest najważniejsza.
Sugestia: Zastosowanie metryk uwzględniających nierównowagi w danych, np. ważony F1-score. Implementacja oceny modelu przy użyciu technik takich jak bootstrap, aby zrozumieć zmienność wydajności modelu.

8 Interpretowalność modelu

Problem: Podstawowe wykorzystanie ważności permutacji bez głębszego badania przyczyn decyzji modelu.
Sugestia: Integracja zaawansowanych narzędzi interpretowalności, takich jak SHAP (SHapley Additive exPlanations) czy LIME (Local Interpretable Model-agnostic Explanations), które dostarczają wglądu w decyzje modelu.

9 Dokumentacja i komentarze

Problem: Niewystarczająca dokumentacja i komentarze w kodzie.
Sugestia: Opracowanie kompleksowej dokumentacji, która obejmuje przegląd projektu, cele, szczegółowe instrukcje konfiguracji, opisy metodologii, w tym przetwarzania danych, budowy modelu i strategii ewaluacji, a także szczegółowe komentarze w kodzie, wyjaśniające cel każdej funkcji i kluczowych bloków kodu.

Wyniki Testów Walidacyjnych

Wyniki uzyskane na danych walidacyjnych prezentują się następująco:

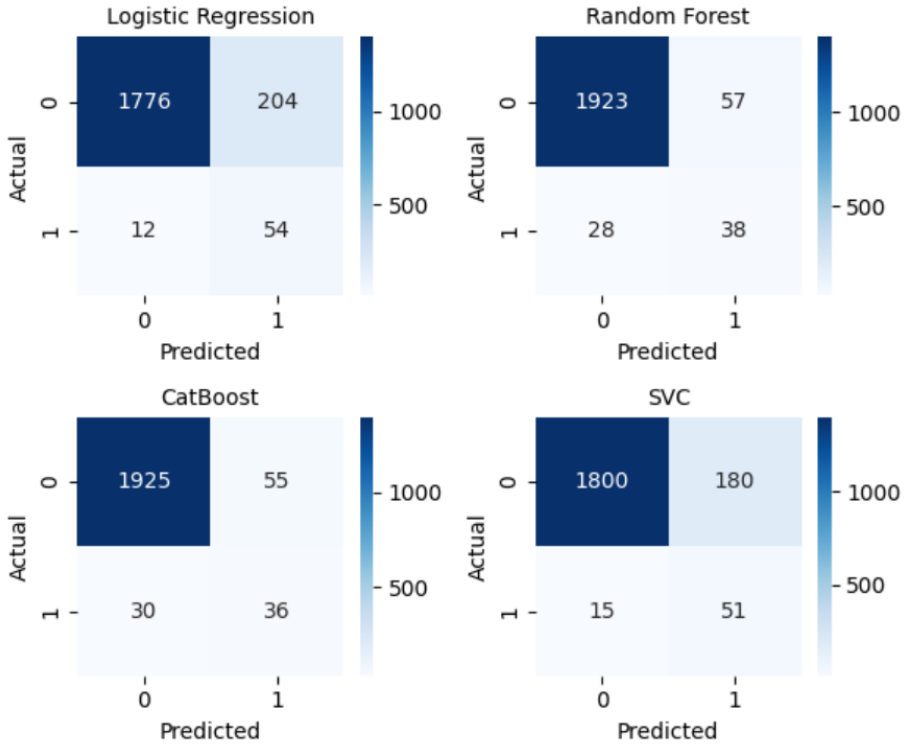
Metric	LogReg	RandFor	Cat
Accuracy	0.89443	0.95846	0.95846
Precision	0.20930	0.40000	0.39560
Recall	0.81818	0.57576	0.54545
F1-Score	0.33333	0.47205	0.48780
AUC-ROC	0.85758	0.77348	0.84091
GINI	0.71515	0.54697	0.68182

Tabela 1: Performance metrics for Logistic Regression, Random Forest, and CatBoost models.

Te wyniki wskazują na pewne ograniczenia modelu, szczególnie w kontekście recall i F1 Score, które są znacząco niższe w porównaniu do accuracy. Niska wartość ROC AUC również sugeruje, że model ma problemy z efektywnym rozróżnianiem klas.

Macierz Błędów

Poniżej przedstawiono macierz błędów (Confusion Matrix), która pozwala na bardziej szczegółowe zrozumienie wydajności modelu:



Rysunek 1: Confussion Matrix

Analiza Confusion Matrix może dostarczyć wglądu w typy błędów popełnianych przez model, co jest kluczowe dla dalszej optymalizacji i zrozumienia, jakie zmiany mogą być konieczne do poprawy ogólnej skuteczności.

Podsumowanie

Implementacja powyższych sugestii znacząco podniesie poziom profesjonalizmu projektu, co jest kluczowe, biorąc pod uwagę jego potencjalny wpływ na życie ludzi w kontekście decyzji finansowych.