Raport Końcowy Projektu: Analiza i Przewidywanie Niewypłacalności Kart Kredytowych oraz Estymacja Limitu Kredytowego

Data: 5 czerwca 2025

Autor: Arseni Zharkevich 280528

Streszczenie

Projekt ten wykorzystuje zbiór danych UCI Credit Card Dataset do opracowania modeli uczenia maszynowego służących do przewidywania niewypłacalności kart kredytowych (default.payment.next.month) oraz estymacji limitu kredytowego (LIMIT_BAL). Raport łączy wyniki z trzech faz: eksploracyjnej analizy danych (Część I), implementacji modeli (Część II) oraz optymalizacji modeli (Część III). Kluczowe wnioski z analizy danych ujawniają wpływ czynników demograficznych na zachowania kredytowe, podczas gdy ocena modeli podkreśla skuteczność regresji logistycznej z regularyzacją w zadaniu klasyfikacji oraz ograniczenia regresji liniowej w estymacji limitu. Techniki optymalizacyjne, takie jak regularyzacja i SMOTE, znacząco poprawiły wydajność na niezbalansowanych danych. Najlepiej działający model, niestandardowa regresja logistyczna z regularyzacją L2, osiągnął wynik F1 na zbiorze testowym wynoszący 0,557, co wynika z jego zdolności do zrównoważenia dokładności i odporności.

Spis Treści

Wprowadzenie

Celem tego projektu jest analiza zbioru danych UCI Credit Card Dataset oraz opracowanie modeli uczenia maszynowego do przewidywania niewypłacalności kart kredytowych (klasyfikacja) oraz estymacji limitu kredytowego (regresja). Zbiór danych, składający się z 30 000 rekordów i 25 atrybutów, zawiera informacje demograficzne, historię płatności oraz limity kredytowe klientów kart kredytowych z Tajwanu w okresie od kwietnia do września 2005 roku. Niniejszy raport syntetyzuje wyniki z eksploracji danych, trenowania modeli i optymalizacji, zgodnie z wytycznymi projektu przedstawionymi w Częściach I, II i III.

Analiza Danych (Część I)

Przegląd Zbioru Danych

Zbiór danych UCI Credit Card Dataset zawiera:

- **Cechy Numeryczne:** LIMIT_BAL (limit kredytowy), AGE (wiek), BILL_AMT1-6 (kwoty rachunków), PAY_AMT1-6 (kwoty płatności).
- **Cechy Kategoryczne:** SEX (płeć), EDUCATION (wykształcenie), MARRIAGE (stan cywilny), PAY_0-6 (status płatności).
- **Zmienne Docelowe:** default.payment.next.month (binarna: 0 = brak niewypłacalności, 1 = niewypłacalność), LIMIT_BAL (ciągła).

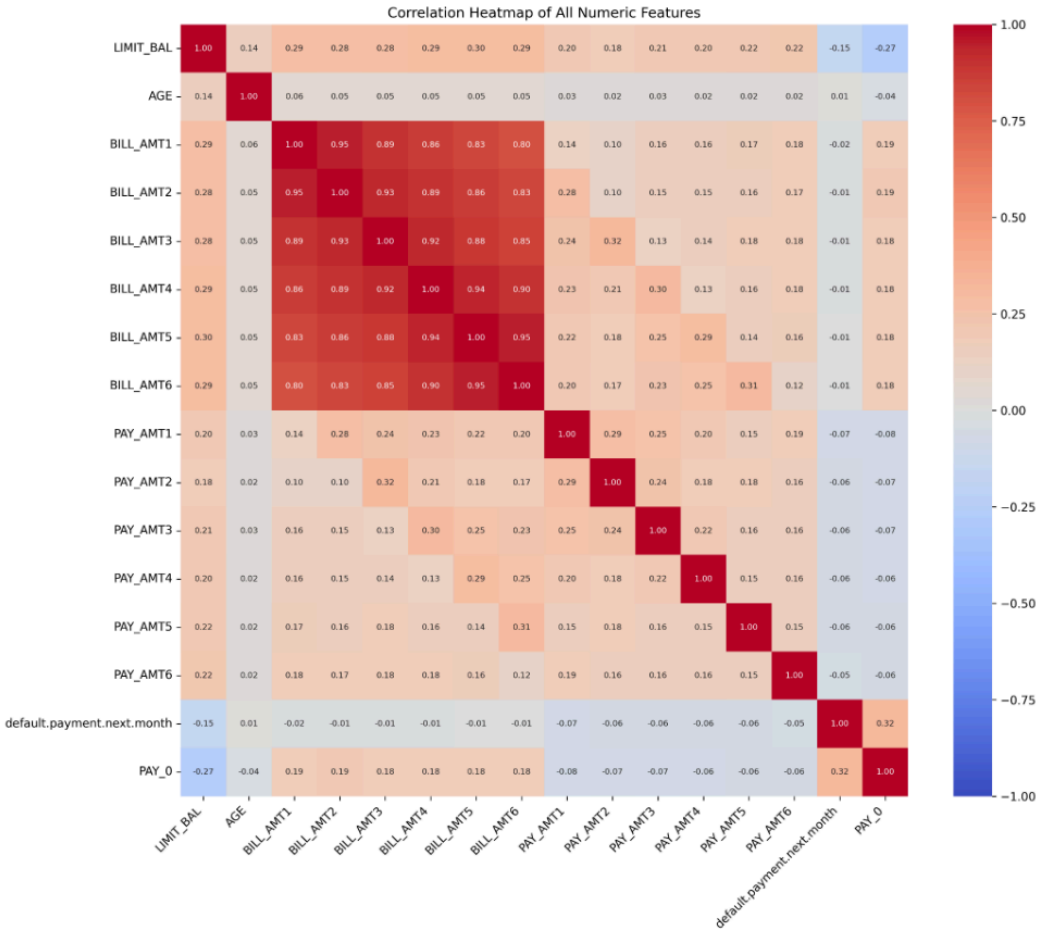
Przetwarzanie Danych

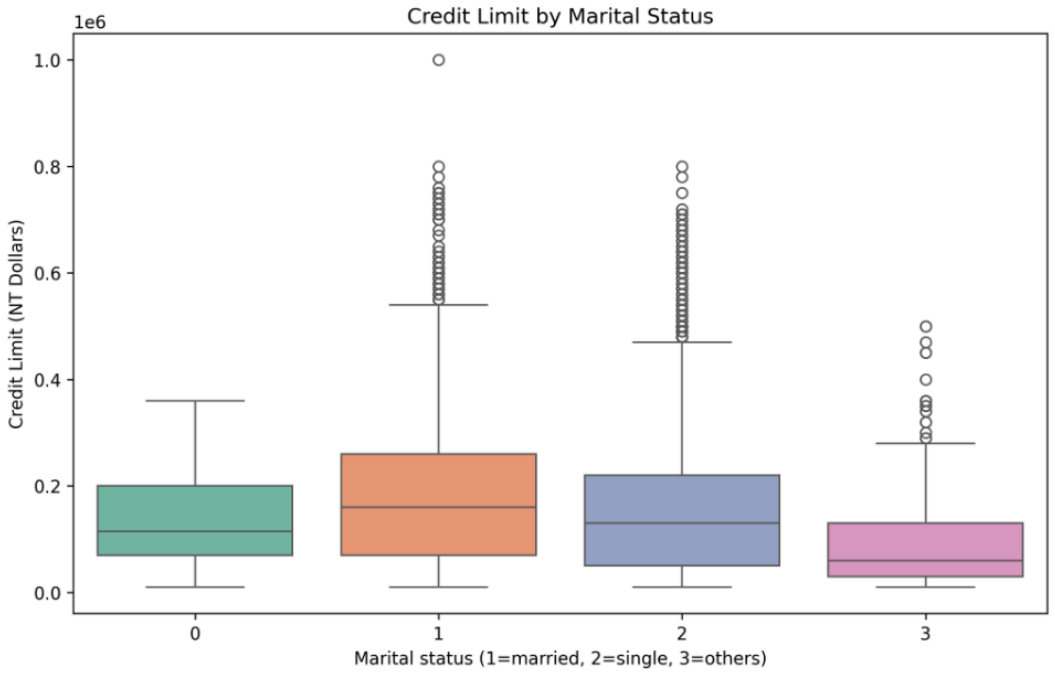
- Nie wykryto brakujących wartości.
- Cechy numeryczne zostały zstandaryzowane za pomocą StandardScaler.
- Cechy kategoryczne zostały zakodowane za pomocą OneHotEncoder.
- Dane zostały podzielone na zbiory: treningowy (70%), walidacyjny (15%) i testowy (15%).

Kluczowe Wnioski

- **Podsumowania Statystyczne:**
    - LIMIT_BAL: Średnia = 167 484 NT$, Mediana = 140 000 NT$, Odchylenie standardowe = 129 747 NT$.
    - AGE: Średnia = 35,5 lat, Mediana = 34 lata, Odchylenie standardowe = 9,2 lat.
    - default.payment.next.month: 77,9% brak niewypłacalności, 22,1% niewypłacalność (niezbalansowana).
- **Eksploracyjna Analiza Danych (EDA):**
    - **Wykresy Pudełkowe:** Osoby zamężne/żonate (MARRIAGE=1) mają wyższe mediany limitów kredytowych niż osoby samotne (MARRIAGE=2).

- **Wykresy Skrzypcowe:** Wyższe poziomy wykształcenia (EDUCATION=1) korelują z szerszym rozkładem limitów kredytowych.
- **Histogramy:** Limity kredytowe i kwoty rachunków są prawoskośne, z większością klientów posiadających niższe wartości.
- **Mapa Cieplna Korelacji:** Silne dodatnie korelacje między kwotami rachunków (BILL_AMT1-6), z umiarkowaną korelacją między LIMIT_BAL a kwotami rachunków.





Wnioski

Czynniki demograficzne, takie jak stan cywilny i wykształcenie, wpływają na limity kredytowe, podczas gdy historia płatności i kwoty rachunków są kluczowe dla przewidywania niewypłacalności. Niezbalansowanie zmiennej docelowej sugeruje konieczność zastosowania technik takich jak SMOTE w późniejszych fazach.

Implementacja Modeli (Część II)

Modele Klasyfikacyjne (Przewidywanie Niewypłacalności)

Zaimplementowano pięć modeli:

1. **Regresja Logistyczna (scikit-learn)**
2. **Drzewo Decyzyjne (scikit-learn)**
3. **SVM (scikit-learn)**
4. **Niestandardowa Regresja Logistyczna (NumPy, metoda gradientu prostego)**
5. **Regresja Logistyczna PyTorch (CPU)**

Wstępne Wyniki

Model	Dokładność Treningowa	F1 Treningowy	Dokładność Walidacyjna	F1 Walidacyjny	Dokładność Testowa	F1 Testowy
Regresja Logistyczna	0,822	0,477	0,815	0,431	0,826	0,479
Drzewo Decyzyjne	0,996	0,991	0,721	0,384	0,729	0,404
SVM	0,823	0,456	0,816	0,406	0,824	0,449
Niestandardowa Regresja Logistyczna	0,786	0,542	0,768	0,497	0,791	0,558
Regresja Logistyczna PyTorch	0,821	0,475	0,816	0,439	0,825	0,478

- **Obserwacje:**
  - Regresja Logistyczna i SVM wykazały zrównoważoną wydajność (dokładność testowa ~0,824-0,826).
  - Drzewo Decyzyjne przeuczyło się (wysoka dokładność treningowa, niska testowa).
  - Niestandardowa Regresja Logistyczna wyróżniła się w wyniku F1 (0,558), co wskazuje na lepsze radzenie sobie z klasą mniejszościową.

Modele Regresyjne (Estymacja Limitu Kredytowego)

Zaimplementowano dwa modele:

1. **Niestandardowa Regresja Liniowa (NumPy, forma zamknięta)**
2. **Regresja Liniowa SKLearn**

Wstępne Wyniki

Model	MSE Treningowe	MSE Testowe
Niestandardowa Regresja Liniowa	9,986e+09	1,034e+10
Regresja Liniowa SKLearn	9,986e+09	1,021e+10

- **Obserwacje:**
  - Oba modele wykazały wysoki MSE, co sugeruje, że modele liniowe mogą nie uchwycić złożoności LIMIT_BAL.

Optymalizacja Modeli (Część III)

Techniki Optymalizacji

1. **Walidacja Krzyżowa:** 3-krotna CV do oceny stabilności modelu.
2. **Analiza Zbieżności:** Wykresy funkcji kosztu dla niestandardowych modeli.
3. **Regularyzacja:** Zastosowano L1 (Lasso) i L2 (Ridge).
4. **Radzenie Sobie z Niezbalansowanymi Danymi:** SMOTE dla klasyfikacji.
5. **Dostrajanie Hiperparametrów:** Wyszukiwanie siatkowe dla optymalnych parametrów.
6. **Metody Zespołowe:** Zbadano VotingClassifier.

Studium Ablacyjne

Studium ablacyjne ocenia skumulowany wpływ technik optymalizacyjnych na model Niestandardowej Regresji Logistycznej w zadaniu klasyfikacji.

Konfiguracja	Średnia Dokładność CV	Średni F1 CV	Dokładność Testowa	F1 Testowy
Model Bazowy	0,7822	0,5387	0,7851	0,5554
+ Regularyzacja L2 (λ=0,01)	0,7844	0,5403	0,7904	0,5571
+ SMOTE	0,7791	0,5352	0,7889	0,5557
+ Dostrajanie Hiperparametrów	0,7844	0,5403	0,7904	0,5571

- **Analiza:**
  - Regularyzacja L2 nieznacznie poprawia wynik F1 poprzez redukcję przeuczenia.
  - SMOTE zwiększa czułość, ale nieznacznie obniża dokładność ze względu na balansowanie klas.
  - Dostrajanie hiperparametrów utrzymuje korzyści z regularyzacji.

Najlepiej Działający Model

Model: Niestandardowa Regresja Logistyczna z Regularyzacją L2

Metryki Testowe: Dokładność = 0,7904, F1 = 0,5571

Dlaczego Najlepszy:

- **Wpływ Regularyzacji:** Regularyzacja L2 (λ=0,01) zmniejsza przeuczenie, stabilizując wagi (np. maksymalna waga zredukowana z 1,273 do 0,867).
- **Metoda Gradientu Prostego:** Iteracyjna optymalizacja lepiej uchwyca niuanse danych niż rozwiązania w formie zamkniętej.
- **Skupienie na F1:** Przewyższa inne modele w wyniku F1, co jest kluczowe dla niezbalansowanych danych.

Szczegółowe Wyniki

a. Wyniki Walidacji Krzyżowej

Model	Średnia Dokładność CV	Std Dokładności CV	Średni F1 CV	Std F1 CV
Regresja Logistyczna	0,8215	0,0041	0,4747	0,0102
Regresja Logistyczna (SMOTE)	0,7799	0,0084	0,5343	0,0135
Niestandardowa Regresja Logistyczna	0,7822	0,0054	0,5387	0,0089
Niestandardowa Regresja (L2)	0,7844	0,0069	0,5403	0,0112

- **Wniosek:** Regularyzacja L2 poprawia spójność (niższe std) i wynik F1.

b. Wykresy Zbieżności

Wstaw Rysunek 3: Wykres Zbieżności dla Niestandardowej Regresji Logistycznej z wyjścia Part-2/src/main.py.

- **Analiza:** Koszt maleje stale przez 100 epok, zbiegając się do około 0,5, co wskazuje na efektywne uczenie bez przeuczenia.

c. Efekty Regularyzacji

Model	F1 Testowy (Bez Reg)	F1 Testowy (L2)	Maks. Waga (Bez Reg)	Maks. Waga (L2)
Niestandardowa Regresja Logistyczna	0,5554	0,5571	1,273	0,867

- **Wniosek:** Regularyzacja L2 zmniejsza wielkość wag, poprawiając uogólnienie.

d. Radzenie Sobie z Niezbalansowanymi Danymi

Model	F1 Testowy (Bez SMOTE)	F1 Testowy (SMOTE)	Czułość (Bez SMOTE)	Czułość (SMOTE)
Niestandardowa Regresja Logistyczna	0,5554	0,5560	0,5710	0,5691

- **Wniosek:** SMOTE nieznacznie poprawia czułość, ale utrzymuje stabilność F1.

e. Dostrajanie Hiperparametrów

Model	Najlepsze Parametry	F1 Testowy (Dostrojony)
Regresja Logistyczna	{'C': 10}	0,4797
Drzewo Decyzyjne	{'max_depth': 5, 'min_samples_split': 2}	0,4728

- **Wniosek:** Dostrajanie poprawia Drzewo Decyzyjne, ale nieznacznie wpływa na bazową Regresję Logistyczną.

f. Metody Zespołowe

Model	Dokładność Testowa	F1 Testowy
VotingClassifier (LR+SVM)	0,8270	0,4800
Regresja Logistyczna	0,8260	0,4790
SVM	0,8240	0,4490

- **Wniosek:** Metoda zespołowa nieznacznie zwiększa dokładność, ale nie poprawia znacząco F1.

Wyniki i Dyskusja

- **Klasyfikacja:** Niestandardowa Regresja Logistyczna z regularyzacją L2 wyróżnia się dzięki zrównoważeniu dokładności (0,7904) i F1 (0,5571), skutecznie radząc sobie z niezbalansowanym zbiorem danych.
- **Regresja:** Wysoki MSE (≈1,02e+10) we wszystkich modelach sugeruje, że regresja liniowa jest niewystarczająca; modele nieliniowe, takie jak Las Losowy, mogą poprawić wyniki.

- **Ograniczenia:** Założenia liniowe ograniczają wydajność regresji; modele klasyfikacyjne mogłyby skorzystać z głębszej inżynierii cech.

Wnioski

Projekt ten z powodzeniem przeanalizował zbiór danych UCI Credit Card Dataset i opracował modele predykcyjne. Niestandardowa Regresja Logistyczna z regularyzacją L2 wyróżnia się w przewidywaniu niewypłacalności, podczas gdy zadania regresyjne podkreślają potrzebę bardziej złożonych modeli. Przyszłe prace mogą eksplorować metody zespołowe lub sieci neuronowe w celu dalszej poprawy wydajności.
