**Raport**

**Autor:** Artem Lytvynenko

**Projekt:** Credit Card Fraud Detection

**1. Opis danych wejściowych**

Projekt wykorzystuje dataset **Credit Card Fraud Detection** z platformy Kaggle, zawierający **284,807 transakcji** kartami kredytowymi z września 2013 roku. Dataset charakteryzuje się:

1. **30 cech wejściowych**: V1-V28 (cechy po transformacji PCA dla zachowania anonimowości), Time (czas od pierwszej transakcji), Amount (kwota transakcji)
2. **Zmienna docelowa**: Class (0 = transakcja prawidłowa, 1 = oszustwo)
3. **Silne niezbalansowanie klas**: 99.83% transakcji prawidłowych (284,315), 0.17% oszustw (492)
4. **Brak wartości null**: dataset kompletny i gotowy do analizy

Ze względu na transformację PCA większości cech, dataset nie zawiera bezpośrednich informacji osobowych, co zapewnia prywatność użytkowników przy zachowaniu wartości analitycznej.

**2. Strategia podziału danych**

Zastosowano **dwukierunkową strategię** podziału danych:

**2.1 Dataset Niezbalansowany (Oryginalny)**

1. **Zbiór treningowy**: 240,000 próbek (84.2%)
2. **Zbiór testowy**: 22,000 próbek (7.7%)
3. **Zbiór walidacyjny**: 22,807 próbek (8.1%)

**2.2 Dataset Zbalansowany (Undersampling)**

1. **Całkowity rozmiar**: 1476 próbki (984 oszustwa + 492 próbki prawidłowe)
2. **Podział proporcjonalny**: 70% trening (1033), 15% test (221), 15% walidacja (221)
3. **Technika balansowania**: Losowe undersampling klasy większościowej

**3. Wstępna analiza danych**

**Opis histogramów:**

Histogramy pokazują rozkłady cech numerycznych (V1-V28, Time, Amount) oraz klasy transakcji („Class”) dla danych o transakcjach kartami kredytowymi. Większość cech ma rozkład skoncentrowany wokół zera, choć niektóre (np. „Amount”) są mocno asymetryczne.

A screenshot of a graph

Description automatically generated

**Problemy:**

* Duża nierównowaga klas („Class”) – bardzo mało przypadków transakcji oszukańczych.
* Silnie skośny rozkład kwot („Amount”).
* Niektóre cechy wymagają normalizacji lub standaryzacji.

**4. Uzasadnienie wyboru technik i modeli**

**4.1 Wybrane Algorytmy**

1. **Logistic Regression** - model bazowy o wysokiej interpretowalności
2. **Random Forest** - ensemble odporny na overfitting, naturalnie obsługuje niezbalansowanie
3. **Gradient Boosting** - sekwencyjna korekta błędów, wysoka wydajność
4. **Linear SVM** - skuteczny w przestrzeniach wysokowymiarowych
5. **Shallow Neural Network** - nieliniowe wzorce, batch normalization

**4.2 Uzasadnienie Wyboru**

1. **Różnorodność podejść**: Porównanie metod liniowych, ensemble i deep learning
2. **Obsługa niezbalansowania**: Parametry class\_weight='balanced' w modelach sklearn
3. **Interpretowalność vs wydajność**: Od prostych modeli liniowych do sieci neuronowych

**4.3 Preprocessing Pipeline**

1. **MinMaxScaler** dla Time (normalizacja 0-1)
2. **RobustScaler** dla Amount (odporność na outliers)

**5. Wyniki testowe i treningowe**

**5.1 Dataset Zbalansowany - Wyniki Walidacyjne**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Class | Precision | Recall | F1-Score | Accuracy | AUROC |
| Logistic Regression | Not Fraud | 0.96 | 0.93 | 0.94 | 0.94 | 0.987 |
|  | Fraud | 0.93 | 0.96 | 0.95 | 0.94 | 0.987 |
| Random Forest | Not Fraud | 0.93 | 0.97 | 0.95 | 0.95 | 0.987 |
|  | Fraud | 0.97 | 0.93 | 0.95 | 0.95 | 0.987 |
| Gradient Boosting | Not Fraud | 0.94 | 0.93 | 0.94 | 0.94 | 0.958 |
|  | Fraud | 0.93 | 0.94 | 0.94 | 0.94 | 0.958 |
| Linear SVC | Not Fraud | 0.96 | 0.93 | 0.94 | 0.94 | N/A |
|  | Fraud | 0.93 | 0.96 | 0.95 | 0.94 | N/A |
| Shallow Neural Net | Not Fraud | 0.96 | 0.90 | 0.93 | 0.93 | 0.985 |
|  | Fraud | 0.91 | 0.96 | 0.93 | 0.93 | 0.985 |

**5.2 Dataset Niezbalansowany - Wyniki Walidacyjne**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Class | Precision | Recall | F1-Score | Accuracy | AUROC |
| Logistic Regression | Not Fraud | 1.00 | 1.00 | 1.00 | 1.00 | 0.990 |
|  | Fraud | 0.83 | 0.56 | 0.67 | 1.00 | 0.990 |
| Random Forest | Not Fraud | 1.00 | 1.00 | 1.00 | 1.00 | 0.971 |
|  | Fraud | 0.88 | 0.78 | 0.82 | 1.00 | 0.971 |
| Gradient Boosting | Not Fraud | 1.00 | 1.00 | 1.00 | 1.00 | 0.833 |
|  | Fraud | 0.67 | 0.67 | 0.67 | 1.00 | 0.833 |
| Linear SVC | Not Fraud | 1.00 | 1.00 | 1.00 | 1.00 | N/A |
|  | Fraud | 0.81 | 0.47 | 0.60 | 1.00 | N/A |
| Shallow Neural Net | Not Fraud | 0.90 | 1.00 | 0.95 | 0.94 | 0.991 |
|  | Fraud | 1.00 | 0.89 | 0.94 | 0.94 | 0.991 |

**5.3 Kluczowe Metryki**

1. **AUROC**: Główna metryka oceny (Area Under ROC Curve)
2. **F1-Score**: Balans między precision i recall dla klasy oszustw
3. **Recall**: Kluczowy w kontekście biznesowym (wykrycie oszustw)

**6. Analiza wyników i wnioski**

**6.1 Kluczowe Obserwacje**

1. Wpływ balansowania danych: Zastosowanie technik balansowania danych nie poprawiło znacząco wyników ze względu na bardzo niską liczbę transakcji oszukańczych w datasetcie.
2. **Najlepsze modele:**

* Neural Network (AUROC: 0.991) – najwyższa moc dyskryminacyjna.
* Logistic Regression (AUROC: 0.990) – dobry balans prostoty i wydajności

**6.2 Ograniczenia i Dalsze Kroki**

**Ograniczenia:**

1. Undersampling redukuje rozmiar datasetu treningowego
2. Brak informacji o kosztach false positive vs false negative
3. Dataset z 2013 roku może nie odzwierciedlać obecnych wzorców oszustw
4. Zbyt mała liczba przypadków oszustw utrudnia efektywne trenowanie mode

**6.3 Wnioski Końcowe**

Projekt pokazał skuteczność różnych algorytmów uczenia maszynowego w wykrywaniu oszustw kartami kredytowymi. Neural Network uzyskała najlepsze wyniki (AUROC: 0.991), Logistic Regression prawie taki samy (AUROC: 0.990). Kluczowe pozostaje zwiększenie liczby przypadków oszustw w danych oraz odpowiedni dobór metryk oceny, uwzględniających realia biznesow