PSZT - Uczenie Maszynowe

Stawczyk Przemysław 293153, Piotr Zmyślony 268833

Contents

1 Opis zagadnienia

1.1 Treść zadania

Przedstawić wyniki analizy zbioru *Bankruptcy*, opisać procedurę eksperymentalną uczenia maszynowego z wykorzystaniem algorytmów *random forest* i *k-najbliższych sąsiadów* oraz opisać wyniki strojenia parametrów powyższych algorytmów.

1.2 Narzędzia

Skrypty oraz algorytm zostały zaimplementowane w Pythonie 3. Wykorzystano biblioteki: *imblearn.over_sampling.SMOTE, sklearn, numpy, matplotlib, scipy.io, impyute*.

2 Opis preprocesingu i modelowania

2.1 Opis danych wejściowych

Jako dane wejściowe posiadamy 5 plików *.arff*, z których każdy zawiera ekonomiczne wskaźniki z systemu EMIS na temat polskich firm i ich klasyfikację względem tego, czy firmy zbankrutowały po *n* latach od roku, w którym zostały zebrane dane. Liczba *n* lat jest różna dla każdego z plików, od 1 do 5, a każda firma opisana jest przez 64 atrybuty, od *X1* do *X64*.

2.2 Analiza zbioru danych

2.2.1 Brakujące dane

Zaczęliśmy od analizy brakujących danych w wierszach. Jak widać w poniższych wynikach w większości zbiorów około połowa wierszy ma brakujące pola.

	1 rok	*2 lata*	*3 lata*	*4 lata*	*5 lat*
długość	7027	10173	10503	9792	5910
pełne wiersze	3194	4088	4885	4769	3031
wiersze wybrakowane	3833	6085	5618	5023	2879

Następnie przeprowadziliśmy analizę rozkładu brakujących danych w kolumnach i wierszach korzystając z biblioteki pythona *missingno*

Figure 1: 1 rok

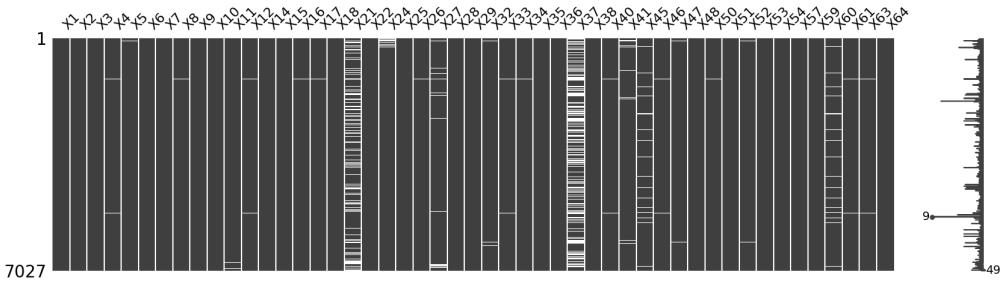


Figure 2: 2 lata

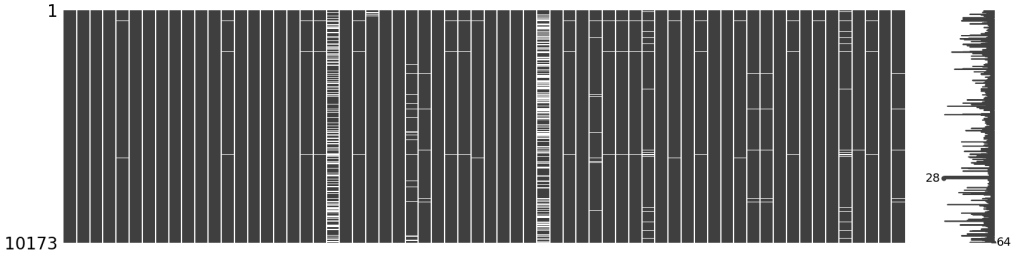
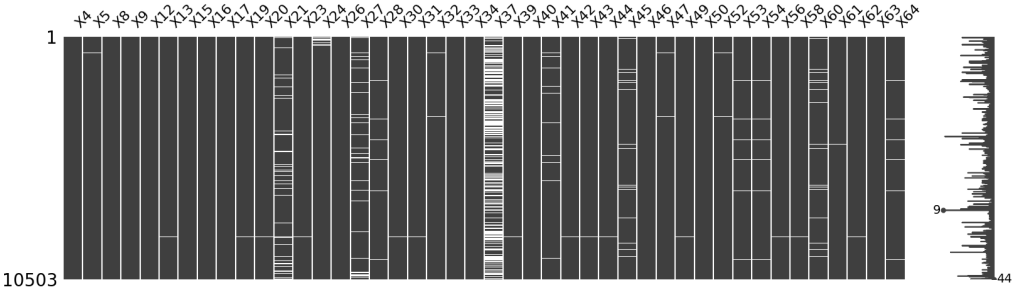


Figure 3: 3 lata



Jak widać większość brakujących danych jest w kolumnie *X37*. Kolumna *X21* ma brakujące w niektórych ale nie wszystkich latach.
Trudno nam było ocenić jaki charakter mają braki w tych danych, czy są skorelowane w wartościami w innych kolumnach czy zupełnie losowe. Aby uniknąć utraty danych dla tych krotek które posiadają wartości w danych kolumnach zdecydowaliśmy się interpolować brakujące dane. W tym celu wybraliśmy 4 metody:

1. Wstawianie średniej w danej kolumnie *(Jako punkt odniesienia)*
2. K najbliższych krotek
3. Spodziewanej Maksymalizacji *(Expected Maximalisation)*
4. Algorytm MICE

2.2.2 Zbalansowanie danych

Dokonaliśmy analizy ile z poszczególnych rekordów należy do klas klasyfikacyjnych

Czy zbankrutowano:	*rok 1*	*rok 2*	*rok 3*	*rok 4*	*rok 5*
Tak	6756	9773	10008	9277	5500
Nie	271	400	495	515	410
procent większości	3.857 %	3.932 %	4.713 %	5.259 %	6.937 %

Dane w zbiorach są mocno niezbalansowane dlatego zdecydowaliśmy się na interpolację korzystając z metody *SMOTE (Synthetic Minority Over Sampling Technique)*

Figure 4: 4 lata

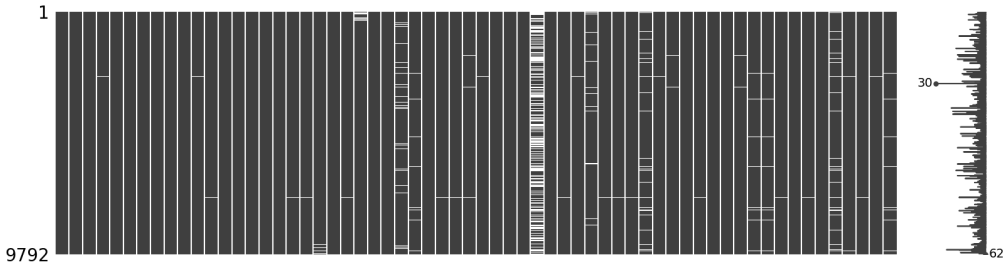
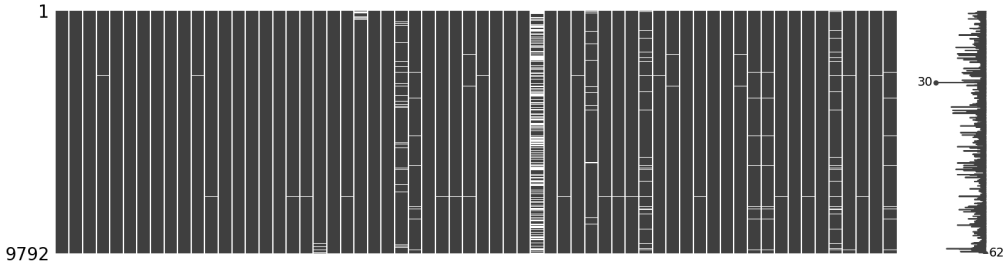


Figure 5: 5 lat



2.3 Przepływ Danych

Po powyższej analizie zdecydowaliśmy o następującym przepływie oryginalnych danych do konstrukcji modeli.
Walidacji modeli planujemy dokonać korzystając K-krotnej walidacji krzyżowej.

2.3.1 Wizualizacja Przepływu Danych

3 Modele

3.1 Parametry Modeli

4 Wyniki Eksperymentu

4.1 Wykresy

4.2 Interpretacja

Figure 6: Przepływ Danych