

PSZT - Uczenie Maszynowe

Stawczyk Przemysław 293153, Piotr Zmyślony 268833

Contents

1	Opis Preprocesingu i Modelowania	2
1.1	Analiza Zbioru danych	2
1.1.1	Brakujące dane	2
1.1.2	Zbalansowanie danych	3
1.2	Przepływ Danych	3
1.2.1	Wizualizacja Przepływu Danych	4
2	Modele	4
2.1	Parametry Modeli	4
3	Wyniki Eksperymentu	4
3.1	Wykresy	4
3.2	Interpretacja	4

1 Opis Preprocesingu i Modelowania

1.1 Analiza Zbioru danych

1.1.1 Brakujące dane

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

	<i>rok 1</i>	<i>rok 2</i>	<i>rok 3</i>	<i>rok 4</i>	<i>rok 5</i>
<i>długość</i>	7027	10173	10503	9792	5910
<i>pełne wiersze</i>	3194	4088	4885	4769	3031
<i>brakujące dane</i>	3833	6085	5618	5023	2879

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

Figure 1: rok 1

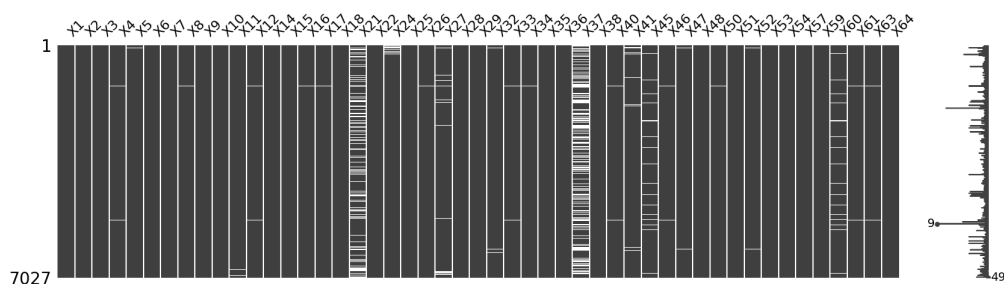


Figure 2: rok 2

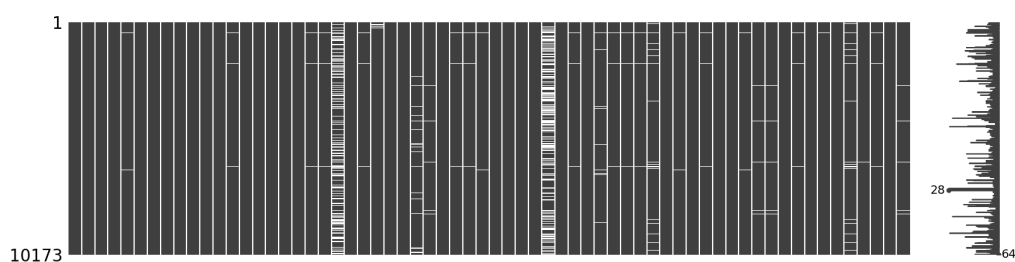


Figure 3: rok 3

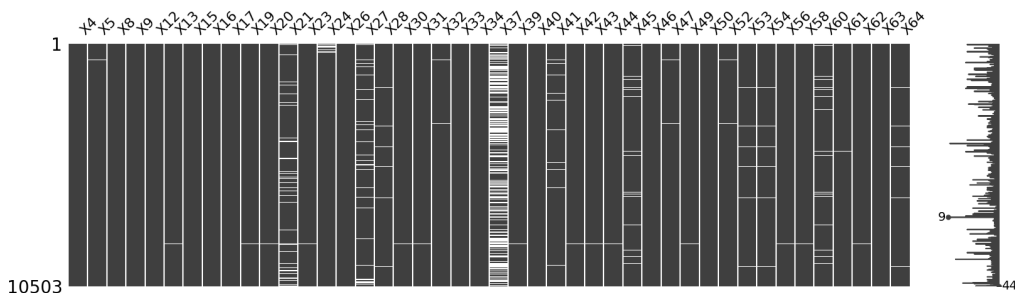


Figure 4: rok 4

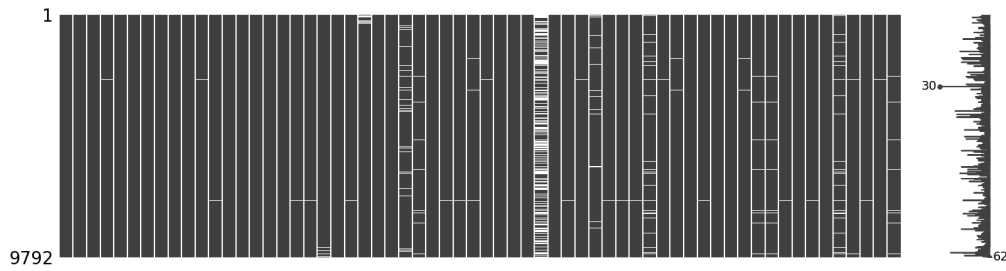
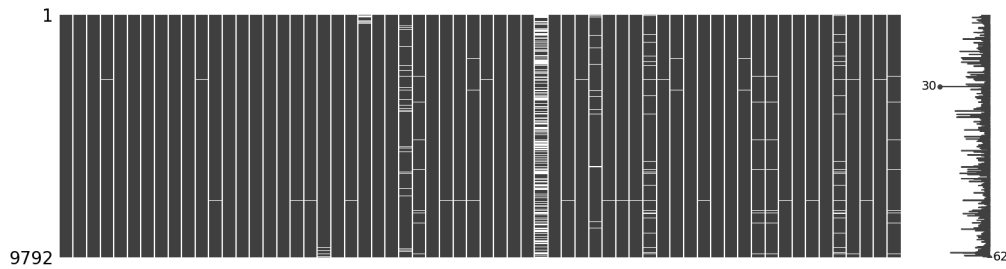


Figure 5: rok 5



Jak widać większość brakujących danych jest w kolumnie X_{37} . Kolumna X_{21} 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

1.1.2 Zbalansowanie danych

Dokonałiśmy analizy ile z poszczególnych rekordów należy do klas klasyfikacyjnych

<i>Czy zbankrutowano:</i>	<i>rok 1</i>	<i>rok 2</i>	<i>rok 3</i>	<i>rok 4</i>	<i>rok 5</i>
<i>Tak</i>	6756	9773	10008	9277	5500
<i>Nie</i>	271	400	495	515	410
<i>procent większości</i>	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*)

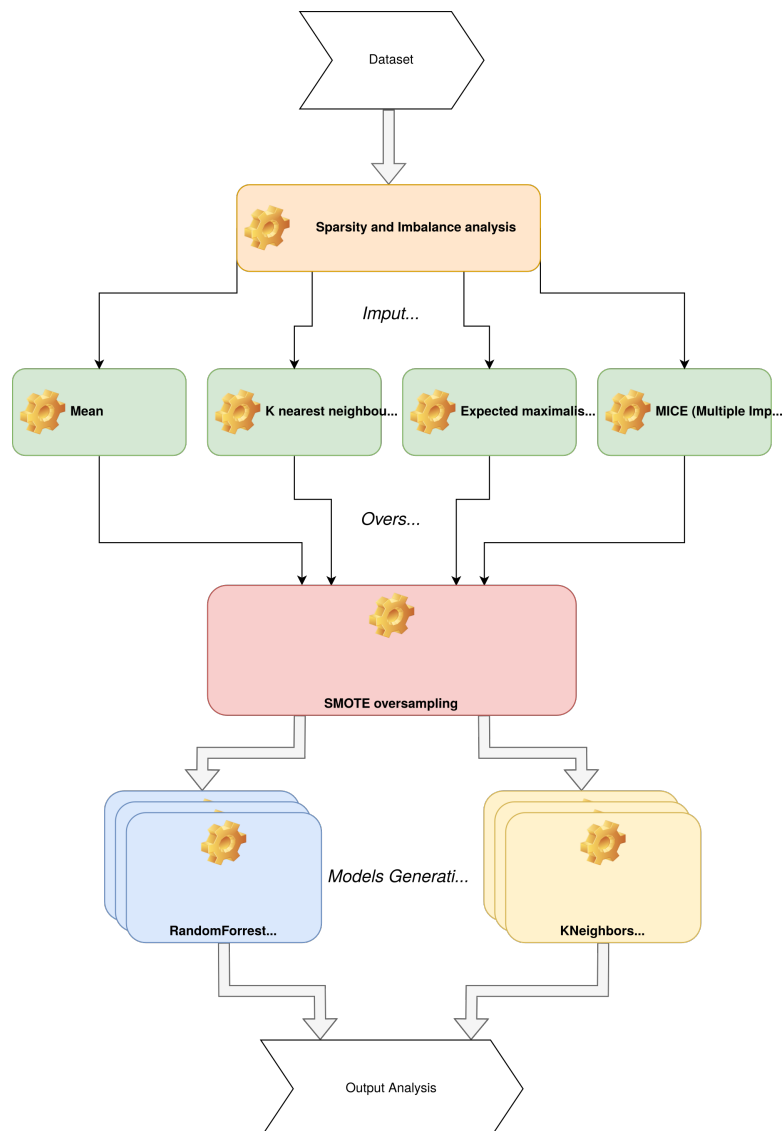
1.2 Przepływ Danych

Po powyższej analizie zdecydowaliśmy o następującym przepływie oryginalnych danych do konstrukcji modeli.

Walidacji modeli planujemy dokonać korzystając K-krotnej walidacji krzyżowej.

1.2.1 Wizualizacja Przepływu Danych

Figure 6: Przepływ Danych



2 Modele

2.1 Parametry Modeli

3 Wyniki Eksperymentu

3.1 Wykresy

3.2 Interpretacja