



# Przewidywanie niewypłacalności kredytobiorców na podstawie danych Credit Score

Pahasian Milanna, Bokhan Katsiaryna, Badzeika Hleb



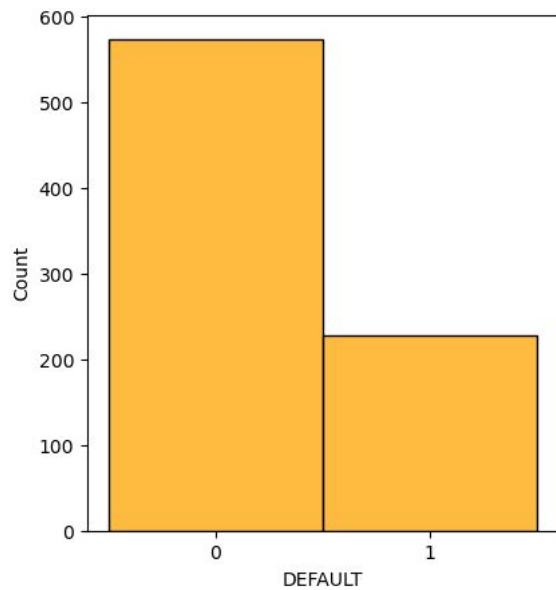
# Dane

87 kolumn, 1000 obserwacji

1. Potencjalne targety: **DEFAULT** lub **CREDIT\_SCORE**
2. Wydatki po czasie (12 i 6 miesiące)
3. Kategorie, npr. Posiadanie karty, kredyt hipoteczny i t.d.
4. Sztucznie wygenerowane dane
5. **Ratios!!**



## Target – 'DEFAULT'

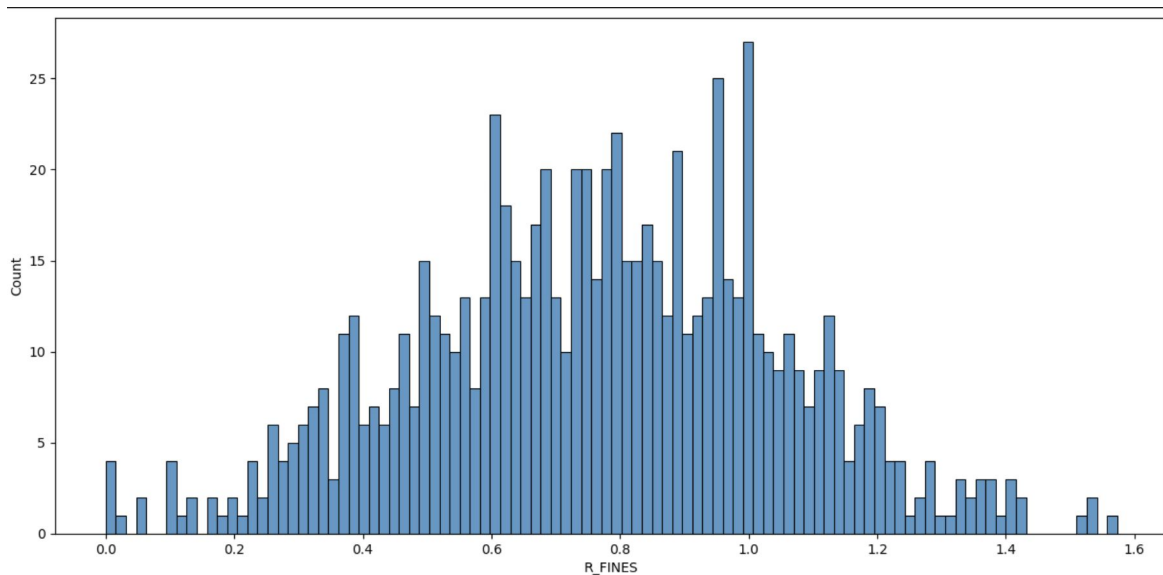


Dataset jest niezbalansowany według zmiennej celu

# Imputacja

Mamy dużo (zastąpionych) NaNów w  
datasetcie, więc wykorzystaliśmy **KNNImputer**  
oraz **IterativeImputer**

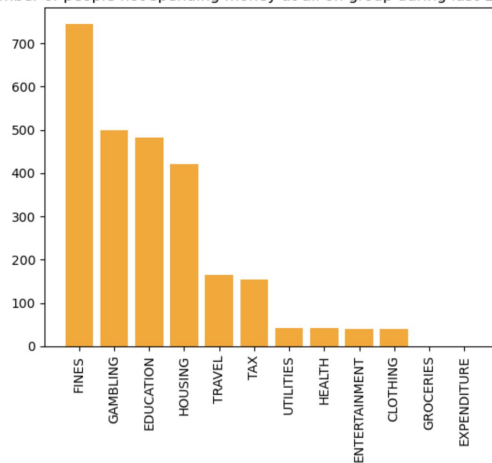
To nie dało wyników



# Praca z danymi

## Liczba zer

Number of people not spending money at all on group during last 12 months



Oraz wyrzuciliśmy outlierów z:

- SAVINGS
- DEBT
- T\_CLOTHING\_12
- T\_CLOTHING\_6
- T\_HEALTH\_12
- T\_HEALTH\_6
- T\_TRAVEL\_12
- T\_TRAVEL\_6



Również postanowiliśmy wyrzucić:

- T\_EXPENDITURE\_12
- T\_EDUCATION\_6
- T\_ENTERTAINMENT\_6
- T\_GAMBLING\_6
- T\_GROCERIES\_6
- R\_UTILITIES\_DEBT
- T\_HOUSING\_6
- T\_TAX\_6
- T\_UTILITIES\_6

Z tego powodu, że te kolumny mają korelacje >95% z innymi kolumnami datasetu

---

# Rozważane modele



## Wyniki

Model	Accuracy (c-v)	Recall 1 (c-v)	Accuracy (T)	Recall 1 (T)
Logistic Regression	0.734375	0.231	0.74375	0.22
Random Forest	0.734	0.127	0.742	0.16
Stacking	0.732	0.143	0.74	0.22
Support Vector Classification	0.731	0.094	0.744	0.11
Gradient Boosting Classifier	0.7297	0.1264	0.73125	0.11
XGBOOST Classifier	0.7234375	0.115	0.71875	0.089













## Logistic Regression Pipeline



Po tym Pipeline'e został użyty GridSearch



## Logistic Regression + SMOTE +Oversampling

Model	Accuracy (c-v)	Recall 1 (c-v)	Accuracy (T)	Recall 1 (T)
Logistic Regression	0.734375	0.231	0.74375	0.22
Logistic Regression + SMOTE	0.774 	0.6179 	0.70625 	0.044 
Logistic Regression + OverSampling	0.627 	0.70 	0.64375 	0.667 

---

# AutoML



## TPOT (Tree-Based Pipeline Optimization Tool)

Logistic  
Regression

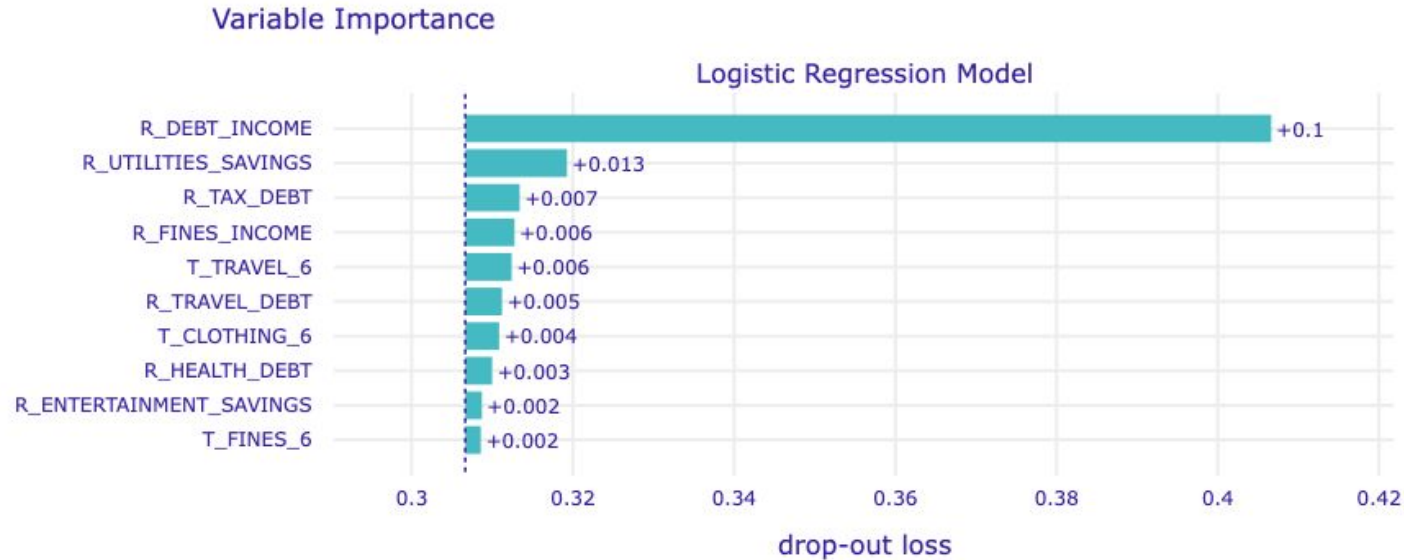


- TPOT:
- 30 gen
  - population size = 40

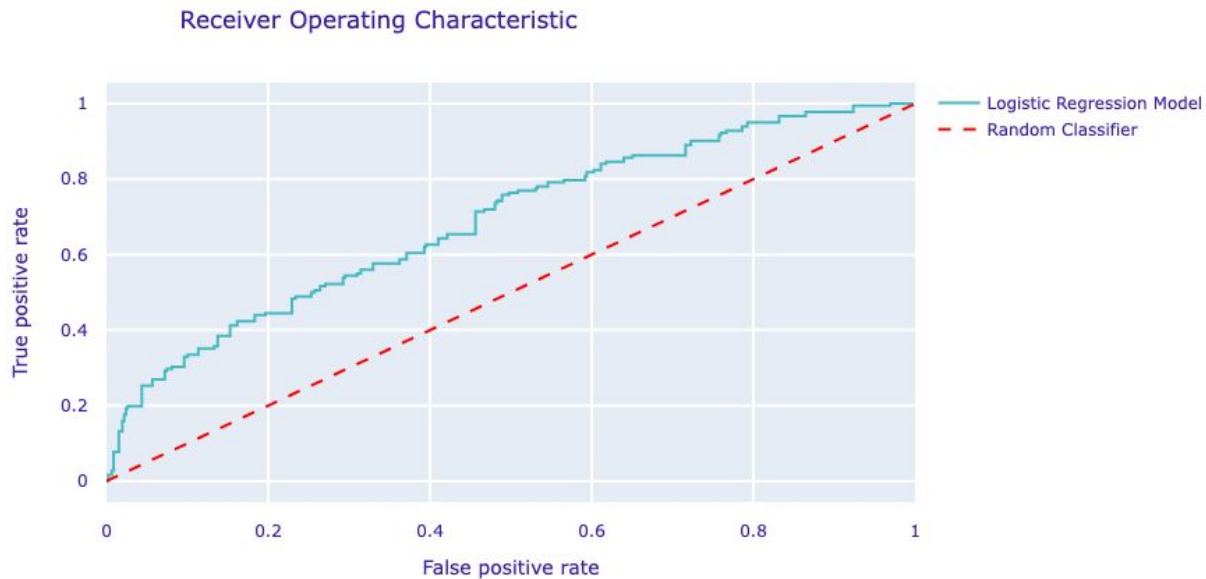
---

xAI

# Feature Importance



# Wydajność modeli



---

# Implikacje biznesowe





## Bank jest zainteresowany:

- Zmniejszenie ryzyka kredytowego
- Zwiększenie dochodowości
- Optimalizacja ofert produktowych



## Co potrafi zbudowany model?

1. Nie potrafi złapać dostateczny procent DEFAULTów ❌
2. Może złapać prawie 95% klientów spłacających kredyt +
3. Daje sensowne i intuicyjne wyniki stosownie ważności zmiennych +



**Dziękujemy za uwagę!**