Raport z realizacji zadania rekrutacyjnego
Stanowisko: Banking Champions – Big Data
Autor: Damian Kos

1. Opis zadania
   Celem zadania było stworzenie modelu modelu uczenia maszynowego przewidującego poważne zaległości w spłacie z wykorzystaniem ogólnodostępnego zbioru danych Home Credit z platformy Kaggle.

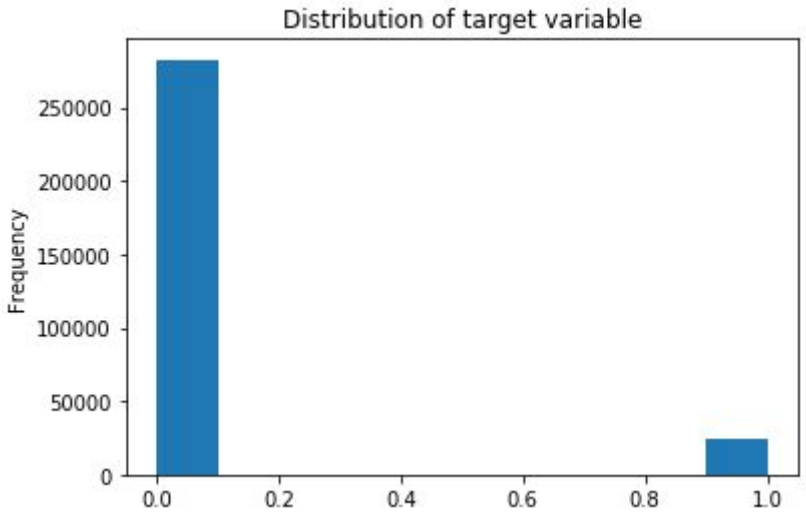
2. Wybór modelu
   Analizę przeprowadziłem z wykorzystaniem metody uczenia maszynowego jaką jest **las losowy (random forest).** Jest to często stosowane podejście w przypadku gdy nie chcemy niepotrzebnie komplikować analizy, a zaprezentować jedynie wstępne wyniki. Z uwagi na ograniczony czas na wykonanie zadanie i chęć zaprezentowania ogólnego podejścia do zagadnienia wydaje się to optymalnym wyborem.

3. Użyte dane
   Opisywany model zbudowany został jedynie na podstawie danych zawartych w plikach **application_{train|test}.csv .** Analiza z wykorzystaniem wszystkich danych dostępnych w ramach konkursu wymagałaby dogłębnego zbadania zagadnienia co z całą pewnością pozwoliłoby na osiągnięcie lepszych rezultatów ale również znacząco wydłużyło czas analizy. Wykorzystanie podstawowego zbioru pozwala natomiast w przystępny sposób zaprezentować ogólne podejście do omawianego problemu.

4. Opis środowiska
   Analiza przeprowadzona została w **Pythonie 3** z wykorzystaniem narzędzia **jupyter notebook.** Całość możliwa jest do odtworzenia za pomocą pliku **Risk_estimation.ipynb**

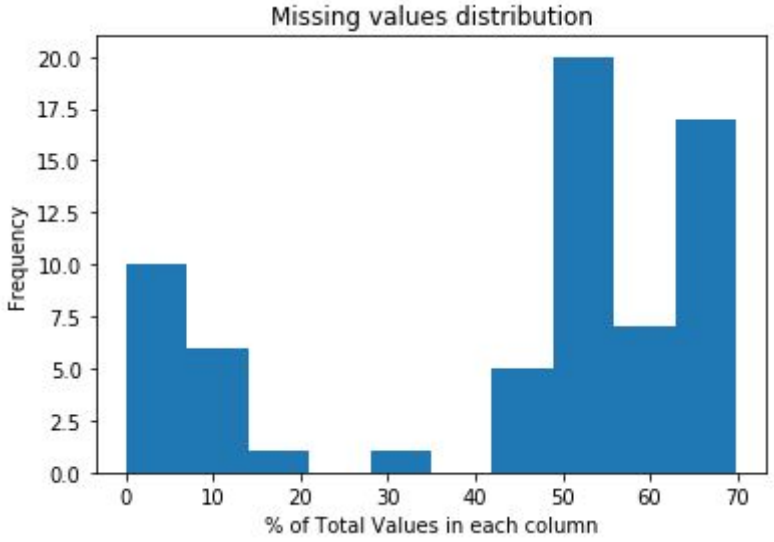
5. Szczegóły implementacji
   Analiza podzielona została na kilka podmodułów:

   5.1 Wczytanie danych (Reading data)
   Dane wczytane zostały z wykorzystaniem biblioteki **pandas** z dwóch plików application_train.csv oraz application_train.csv. Zbiór treningowy zawiera 307511 obserwacji natomiast testowy 48744. W obu tych zbiorach każda obserwacja składa się z 121 zmiennych, a w zbiorze treningowych dodana jest kolumna TARGET określająca czy w przypadku danego kredytu wystąpiły poważne zaległości w spłacie. Warto zwrócić uwagę na rozkład zmiennej objaśnianej. Na podstawie poniższego histogramu widzimy, że zaległości w spłacie obserwowane są stosunkowo rzadko.



5.2 Analiza braków w danych (Missing values)

W trakcie analizy widzimy, że aż 67 z 122 kolumn zawiera pewne braki w danych. Natomiast dokładny rozkład tego jaką część wszystkich obserwacji w danej kolumnie stanowią braki w danych przedstawia poniższy histogram.



Możemy wyróżnić tutaj dwie grupy: kolumny dla których braki danych są nieznaczne i stanowią poniżej 40-50% oraz te gdzie brakuje ponad połowy obserwacji. W trakcie analizy zdecydowałem się na użycie jedynie zmiennych objaśniających dla których braki w danych stanowią **mniej niż 50%.** Nie znamy szczegółów odnośnie sposobu zbierania danych przez co użycie danych dla których obserwujemy tak duże braki może negatywnie wpłynąć na zdolności generalizacyjne i odzwierciedlać jedynie zależności w danej próbie, a nie dla ogółu.

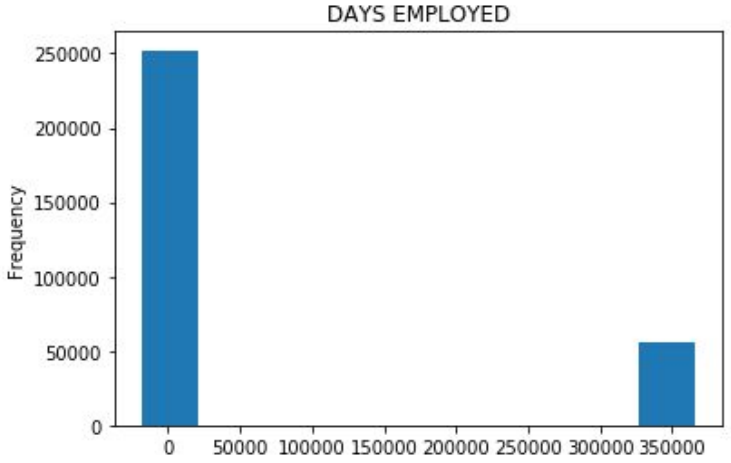
5.3 Transformacja zmiennych (Encoding Categorical Variables)

Przed przystąpieniem do zastosowaniem metod uczenia maszynowego należy dokonać transformacji zmiennych jakościowych. Dla zmiennych w przypadku, których obserwujemy jedynie dwie różne wartości, zastosowałem transformację określaną jako

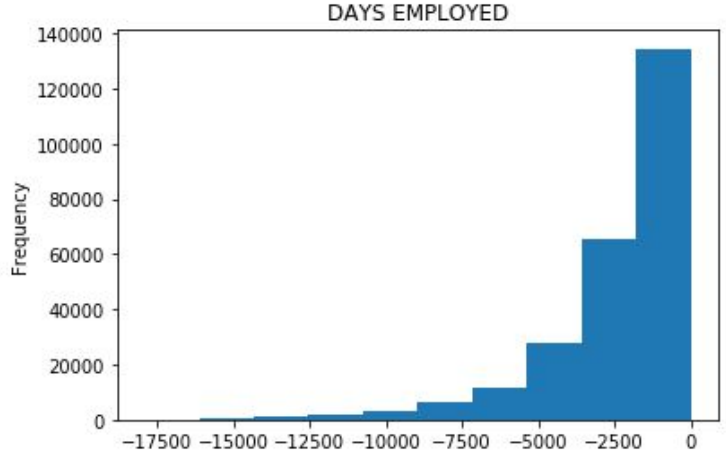
label encoding czyli zamiana na zmienną binarną. Natomiast dla zmiennych przyjmujących więcej niż dwie wartości zastosowałem transformację **one-hot encoding** czyli stworzenie oddzielnej zmiennej binarnej dla każdej z wartości. W przypadku zastosowania podejścia polegającego na przypisaniu kolejnych liczb całkowitych do danych wartości występuje problem kolejności tego przypisania, co może z kolei prowadzić do błędnego określenia zależności przez dany algorytm uczenia maszynowego.

5.4 Analiza danych (Data Exploration)

W tym punkcie sprawdziłem jedynie rozkład tych zmiennych objaśniających, których wartości są łatwe do zinterpretowania. W trakcie analizy odkryłem pewną anomalię występująca dla zmiennej **DAYS_EMPLOYED,** która to przyjmuje wartość 365243 dla 55374 obserwacji. Jest to z całą pewnością pewien błąd w danych gdyż wartości powinny być wyrażone w liczbach ujemnych określających liczbę dni. Pozostawienie tych błędów znacznie wpływa na rozkład zmiennej (poniższy histogram) i może prowadzić do błędnego działania modelu.



Wszystkie wartości zamieniłem na braki w danych oraz dodałem nową zmienną określająca czy dla danej obserwacji wystąpiła przedstawiona anomalia, gdyż to także może stanowić pewną informację w kontekście przyszłej estymacji modelu. Rozkład zmiennej po opisanej zmianie:



5.5 Uzupełnienie braków w danych i normalizacja zmiennych (Filling missing values and normalization)

Braki w danych uzupełnione zostały przez **medianę obserwacji** z danej kolumny. Natomiast aby zapewnić jednakowy wpływ każdej z zmiennych objaśniających na działanie modelu wszystkie zmienne przeskalowane zostały na wartości z zakresu [0,1].

5.5 Budowa modelu lasu losowego (Random Forest)

W trakcie analizy eksperymentalnie dobrałem wartości następnych parametrów

- liczba drzew użyta do budowy lasu
  Zgodnie z literaturą przedmiotu parametr ustawiłem na wartość odpowiadającą pierwiastkowi kwadratowemu z liczby obserwacji użytych do uczenia modelu, co odpowiada wartości **554**

- liczba obserwacji losowana do budowy każdego z drzew
  Idea lasu losowego polega na budowie możliwie różnych drzew co możemy zapewnić między innymi przez użycie jedynie cześci obserwacji w trakcie budowy każdego z drzew. Najlepsze rezultaty osiągamy w przypadku wyboru około **połowy** obserwacji, a dokładnie 138500.

- maksymalna głębokość każdego z budowanych drzew
  Każde z drzew wchodzących w las losowy powinno być możliwe proste, a jednocześnie opisywać kluczowe zależności widoczne w danych. Parametr ustawiłem na wartość **100,** co przy wszystkich 187 atrybutach stanowi ponad połowę.

6. Wyniki i wnioski

Opisany model pozwolił na osiągnięcie wyniku na poziomie **0,71** co biorąc pod uwagę wynik wygrywający konkurs (0,80) stanowi dobrą podstawę do dalszego rozszerzenia analizy. Przyszłe prace powinny skupiać na poniższych aspektach:

- użycie wszystkich dostępnych danych
- dokładniejsze zbadanie zależności występujących w danych i wykorzystanie techniki zwanej jako **Feature Engineering** czyli wprowadzenie nowych zmiennych uzyskanych na podstawie surowych danych
- zastosowanie efektywniejszych metod uczenia maszynowego ze szczególnym uwzględnieniem **wzmacniania gradientowego (Gradient Boosting)**

Jednak ze względu na poglądowy charakter przeprowadzanej analizy otrzymany wynik traktuje jako satysfakcjonujący.