1. Cel projektu

Tematem naszego projektu będzie prognozowanie bankructwa przedsiębiorstw w Polsce na podstawie ich wskaźników finansowych dotyczących działalności, płynności, wypłacalności i rentowności. Zestaw danych na których będziemy budować swój model to ogólnodostępna baza Polish companies bankruptcy data, w której upadłe firmy były analizowane w latach 2000-2012, natomiast firmy nadal działające były oceniane w latach 2007-2013.

Do stworzenia modelu wykorzystamy kilka algorytmów klasyfikacji binarnej i wybierzemy najlepszy z nich. Będą to:

- Drzewa klasyfikacyjne
- Las losowy
- Sieci neuronowe
- Klasyfikator Naiwny Bayesa
- Klasyfikator XGBoost
- Model logitowy
- Model wykorzystujący sztuczną sieć neuronową

Po budowie modelu stworzymy interfejs użytkownika w Pythonie wykorzystując ipywidgets, w którym dana osoba będzie mogła sprawdzić jakie jest prawdopodobieństwo, że dana firma upadnie w ciągu najbliższego roku. Efektem końcowym naszego projektu będzie prototyp prostego narzędzia weryfikującego płynność finansową przedsiębiorstwa- narzędzie takie wydaje się szczególnie użyteczne w kontekście oceniania potencjalnych kontrahentów na podstawie publicznie dostępnych danych możliwe (możliwość odmówienia realizacji zlecenia na rzecz kontrahenta, który zostanie wytypowany jako zagrożony upadłością, lub obniżenie limitów zamówień jakie taki podmiot może złożyć).

Jako że zestaw danych, na których będziemy pracować zawiera ponad 60 wskaźników naszym celem będzie zminimalizować liczbę zmiennych w modelach, aby końcowi użytkownicy mogli użyć naszej aplikacji bez wpisywania kilkudziesięciu wartości.

Wykorzystując diagnostykę lokalną, oprócz konkretnego prawdopodobieństwa bankructwa będziemy mogli również zapewnić rekomendacje, na którym ze wskaźników (lub na jakiej grupie) należałoby się skupić aby znacznie obniżyć prawdopodobieństwo bankructwa i poprawić bilans aktywów przedsiębiorstwa.

2. Metodyka prowadzenia projektu

Jako że, głównym celem projektu jest dostarczenie wydajnych modeli analitycznych i jego integracja z interfejsem, należy założyć, że głównym wyzwaniem będzie ewaluacja modeli. Po wstępnym wdrażaniu modeli, będą następowały cykliczne fazy:

- optymalizacji hiperparametrów modeli i redukcji wymiarowości modeli

- implementacji modeli dla istniejących danych

- ewentualna faza ETL według potrzeb.

Występująca w procesie cykliczność ewaluacji jest charakterystyczna dla modelu spiralnego- wobec powyższego **właśnie metodyka oparta o model spiralny** zdaje się być właściwym wyborem dla projektu tego typu.

3. Lista członków zespołu:

- **Jadwiga Szkatuła** - Data Analyst - przygotowanie danych do obróbki, konieczne transformacje, budowa modeli i manipulacja hiperparametrami , Zarządzanie pracami nad projektem
- **Krzysztof Wenda** - Data Analyst, zestawienie dopasowania modeli, budowa sieci neuronowej i optymalizacja hiperparametrów sieci neuronowej
- **Filip Prekiel** - Wizualizacje, Interfejs Użytkownika, praca nad intuicyjnością i praktycznoscią rozwiązania

4. Wskaźniki, co mogłoby być miernikiem postępu prac w projekcie.

W trakcie prac będziemy realizować po kolei poszczególne elementy:

1. Analiza opisowa zbioru danych
2. Zarządzanie brakami danych
3. Eliminacja obserwacji odstających
4. Budowa modeli klasyfikacji binarnej (+ manipulacja hiper parametrami)
5. Wybór najlepszego modelu
6. Konstrukcja interfejsu użytkownika

Ostatecznym, wymiernym celem projektu będzie budowa modelu, w którego dokładność będzie wynosiła co najmniej 70%. Jako że, w ramach projektu następowała będzie

ewaluacja modeli, za odpowiednią metrykę stanowiąca miarę postępu prac projektowych w modelu spiralnym należy uznać zwiększającą się dokładność modeli- miara zwiększania dokładności modeli będzie świadczyła o postępie prac projektowych.