Algorytmiczna stabilizacja procesu sprzedaży: transformacja danych w przemyśle maszynowym

W sektorze produkcji zaawansowanych maszyn dla przemysłu spożywczego – od linii pakujących po precyzyjne systemy dozowania – proces sprzedaży rzadko jest szybki. Transakcje są na wysokie kwoty, a cykl decyzyjny jest złożony. W tym środowisku call center przestaje pełnić funkcję prostego biura obsługi, stając się strategicznym hubem, w którym inżynieria spotyka się z negocjacjami.

Celem niniejszego opracowania jest przedstawienie ewolucji systemu analitycznego, który ma na celu optymalizację doboru par Klient-Handlowiec (*Smart Routing*). Źle dopasowany konsultant może zaprzepaścić szansę na kontrakt wart setki tysięcy złotych, podczas gdy optymalne dopasowanie profilu psychologicznego i kompetencyjnego znacząco zwiększa konwersję.

Faza 1: digitalizacja i strukturyzacja głosu

Pierwszym wyzwaniem jest przekształcenie tysięcy godzin rozmów telefonicznych w dane możliwe do przetworzenia. Skala operacji wymaga pełnej automatyzacji.

1. **Transkrypcja** – wykorzystujemy silnik Whisper-1 do zamiany mowy na tekst.

2. **Identyfikacja mówców (diarizacja)** – jest to etap krytyczny. Ponieważ systemy nagrywające często dostarczają jedną ścieżkę audio, konieczne jest rozróżnienie, kto mówi. Wdrożono tu zaawansowany model wnioskujący (najlepszy do tego jest obecnie 01), który na podstawie kontekstu, np. używania pytań ofertowych vs. pytań o specyfikację techniczną, precyzyjnie taguje wypowiedzi jako „Klient" lub „Doradca Techniczny".

Faza 2: ekstrakcja cech psychometrycznych i behawioralnych

Zamiast polegać na intuicji, system dokonuje głębokiej analizy każdej interakcji budując wielowymiarowy wektor cech dla obu stron rozmowy.

- **Profile osobowości** – analiza języka pod kątem modelu Wielkiej Piątki (OCEAN) oraz modelu HEXACO. Każda cecha (np. ekstrawersja, sumienność, uczciwość) jest kwantyfikowana w skali 0-100.
- **Techniki perswazji** – algorytm używa 7 zasad wywierania wpływu wg Cialdiniego (kodowanie binarne 0/1), sprawdzając czy handlowiec stosuje np. regułę niedostępności lub autorytetu i jak reaguje na to klient. Sprawdza też czy klienci używają tych metod.
- **Heurystyki branżowe** – zmienne specyficzne dla sprzedaży maszyn, np.: „poziom wiedzy technicznej klienta", „fiksacja na cenie", „pilność wdrożenia", „skłonność do ryzyka technologicznego".

Faza 3: ograniczenia podejścia archetypowego

Wstępne próby kategoryzacji opierały się na przypisywaniu rozmówców do czterech zdefiniowanych przez LLM archetypów (np. „Inżynier-Formalista", „Wizjoner", „Negocjator"). Choć model ten był łatwy do zrozumienia dla zarządu, okazał się empirycznie niestabilny. Macierze sukcesu oparte na tych etykietach wykazywały dużą zmienność w czasie (brak powtarzalności wyników tydzień do tygodnia), co uniemożliwiało budowę wiarygodnego systemu routingu.

Faza 4: obecny standard – redukcja wymiaru i klasteryzacja (PCA + K-means)

W odpowiedzi na niestabilność archetypów, wdrożono podejście oparte na twardej statystyce (*Baseline*). Jest to metoda nienadzorowana (*unsupervised*), która szuka naturalnych skupisk w danych bez wiedzy o wyniku sprzedaży.

1. **Analiza głównych składowych (PCA)** – ze względu na dużą liczbę zmiennych (ok. 50 cech psychologicznych i behawioralnych) stosuje się PCA do kompresji danych. Pozwala to wyeliminować szum i skorelowane zmienne, redukując przestrzeń do kilku głównych wymiarów.

2. **Algorytm K-means** – na zredukowanej przestrzeni uruchamiany jest algorytm k-średnich. Liczba klastrów (*k*) jest dobierana metodą „łokcia", co pozwala na matematyczne wyznaczenie optymalnej liczby grup klientów i handlowców.

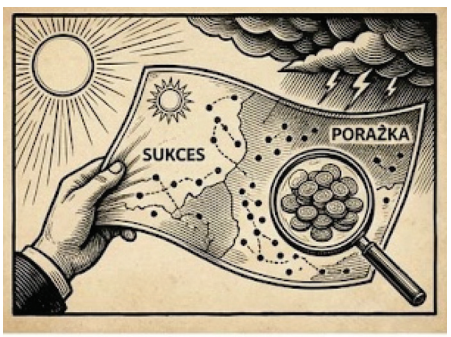
Wady: metoda ta grupuje ludzi podobnych do siebie, ale niekoniecznie tych, którzy generują sprzedaż. Klaster może być spójny psychologicznie, ale niejednorodny pod względem konwersji, co nadal generuje pewien „szum" w macierzy decyzyjnej.

Faza 5: przyszłość – metody alternatywne (*learning to match*)

Aby wyeliminować wady podejścia nienadzorowanego i bezpośrednio powiązać profilowanie z sukcesem sprzedażowym (zmienna celu: 0-1), proponujemy wdrożenie 5 zaawansowanych metod analitycznych.

Alternatywa 1: topologiczne mapowanie sukcesu (supervised UMAP + HDBSCAN)

Zamiast liniowego rzutowania danych (PCA), używamy nieliniowej redukcji wymiarowości UMAP w trybie nadzorowanym.



Ryc. 1. Topologiczne Mapowanie Sukcesu (Supervised UMAP + HDBSCAN). Rozciąganie przestrzeni danych, by oddzielić strefy sukcesu i znaleźć gęste skupiska.

Mechanizm: algorytm otrzymuje informację o tym czy rozmowa zakończyła się sprzedażą. „Rozciąga" przestrzeń cech tak, aby oddzielić przypadki sukcesu od porażki, a dopiero potem grupuje dane.

Zaleta: klastry wynikowe są z definicji silnie skorelowane ze skutecznością sprzedaży. Dodatkowo, użycie algorytmu gęstościowego HDBSCAN pozwala odrzucić „szum" (nietypowe rozmowy), co zwiększa czystość danych co przekłada się na stabilność wyników.

Alternatywa 2: embeddingi drzewiaste (*random forest proximity*)

Wykorzystanie modeli zespołowych (Random Forest/XGBoost) do stworzenia nowej metryki odległości.

Ryc. 2. Embeddingi Drzewiaste (Random Forest Proximity). Ścieżki w lesie decyzyjnym prowadzą podobne przypadki do jednego celu.

Mechanizm: trenujemy model klasyfikacji „Sprzedaż Tak/Nie". Nie używamy go jednak do prognozy, lecz do budowy macierzy podobieństwa. Dwóch klientów jest „bliskich", jeśli w tysiącach drzew decyzyjnych trafiają do tych samych węzłów końcowych.

Zaleta: metoda ta doskonale radzi sobie z nieliniowymi zależnościami (np. „Klient techniczny kupuje tylko u Handlowca-Eksperta, chyba że handlowiec jest bardzo pewny siebie") i jest bardzo stabilna w czasie.

Alternatywa 3: syjamskie sieci neuronowe (*deep metric learning*)

Podejście oparte na uczeniu głębokim (*deep learning*), inspirowane systemami rozpoznawania twarzy.

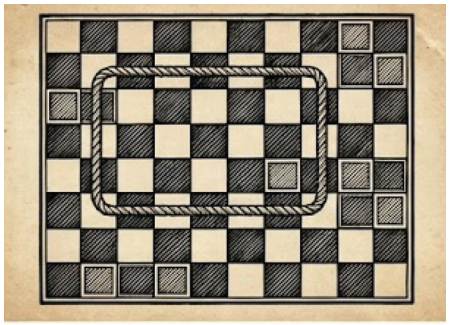
Mechanizm: budujemy sieć neuronową typu Siamese, która uczy się reprezentacji wektorowej (embeddingu) handlowca i klienta. Sieć jest trenowana tak, aby minimalizować odległość między parami, które zawarły transakcję i maksymalizować dla par nieskutecznych.

Ryc. 3. Syjamskie Sieci Neuronowe (Deep Metric Learning). Dwie wieże uczą się wspólnej miary dopasowania, ważąc ważąc sukces i porażkę.

Zaleta: tworzymy dedykowaną „metrykę dopasowania biznesowego". System sam uczy się, jakie kombinacje cech (np. ekstrawersja handlowca + sumienność klienta) prowadzą do sukcesu.

Alternatywa 4: *biclustering* (ko-klasteryzacja widmowa)

Odejście od niezależnego grupowania klientów i handlowców na rzecz jednoczesnego poszukiwania wzorców w macierzy.



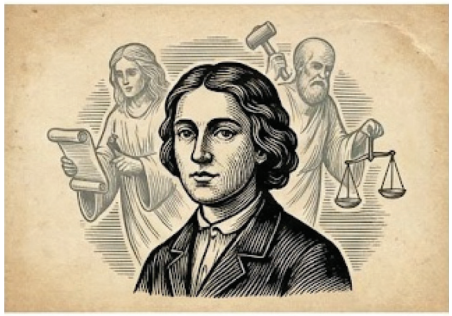
Ryc. 4. Biclustering (Ko-klasteryzacja Widmowa). Odnajdywanie powiązanych bloków klientów i handlowców w macierzy danych.

Mechanizm: algorytm szuka „bloków" w danych – podgrup klientów, które reagują pozytywnie wyłącznie na specyficzne podgrupy handlowców, a negatywnie na inne.

Zaleta: pozwala to wykryć niszowe zależności (lokalne korelacje), które giną w globalnym uśrednianiu metody K-means. Jest to idealne do wyłapywania specyficznych segmentów rynku maszynowego (np. „Klienci innowatorzy" vs „Konserwatywne zakłady produkcyjne").

Alternatywa 5: probabilistyka ukryta (*latent class analysis*/GMM)

Zastosowanie modeli mieszanych (*gaussian mixture models*) zamiast twardego przypisania do grup.



Ryc. 5. Probabilistyka Ukryta (LCA / GMM). Każda osoba jest mieszanką różnych, ukrytych stylów z określonym prawdopodobieństwem.

Mechanizm: model zakłada, że każdy handlowiec i klient jest „mieszanką" różnych stylów z określonym prawdopodobieństwem (np. 80% doradca, 20% sprzedawca).

Zaleta: kluczowa dla stabilizacji wyników w czasie. Jeśli styl handlowca ewoluuje nieznacznie z tygodnia na tydzień, model probabilistyczny koryguje to płynnie. W twardym K-means mała zmiana mogłaby przerzucić handlowca do innej grupy drastycznie zmieniając statystyki macierzy. GMM zapobiega takim skokom.

Podsumowanie i wnioski operacyjne

Przejście z etapu eksperymentalnego (Archetypy) do inżynieryjnego (PCA+K-means, a docelowo metody nadzorowane) jest niezbędne, aby system wsparcia sprzedaży w branży maszynowej stał się narzędziem przewidywalnym.

Rekomenduje się rozpoczęcie testów A/B od metody supervised UMAP lub embeddingów drzewiastych, ponieważ oferują one najlepszy balans między precyzją dopasowania a odpornością na zmienność danych w czasie. Celem nadrzędnym pozostaje zbudowanie stabilnej macierzy prawdopodobieństwa, która pozwoli na automatyczne kierowanie połączeń (routing) z minimalnym ryzykiem błędu.

Wojciech Moszczyński

Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu, specjalista z zakresu ekonometrii, finansów, data science oraz rachunkowości zarządczej. Specjalizuje się w optymalizacji procesów produkcyjnych i logistycznych. Prowadzi badania w obszarze rozwoju i zastosowania sztucznej inteligencji. Od lat zaangażowany w popularyzację machine learning oraz data science w środowiskach biznesowych.