Projekt

Zespół

Skład

Kamil Pietrzyk, Jadwiga Szkatuła, Maciej Śmiałowski, Mikołaj Zapalski

Doświadczenie

Wszyscy członkowie zespołu posiadają wykształcenie wyższe, ukończony pierwszy stopień studiów Informatyka i Ekonometria na Wydziale Zarządzania Akademii Górniczo-Hutniczej.

Dlaczego my (razem)?

Każda z osób w zespole wyróżnia się innym zestawem umiejętności które w połączeniu tworzą niezwykle skuteczną broń w radzeniu sobie z różnymi problemami czy zagadnieniami analitycznymi.

Zapalski Mikołaj - lead developer
Pietrzyk Kamil - data analyst
Jadwiga Szkatuła - debugger, data analyst
Śmiałowski Maciej - data visualization

Problem

Jaki problem klienta chcemy rozwiązać?

Zadaniem, które zostanie rozwiązane w ramach projektu będzie predykcja cen mieszkań w poszczególnych miastach Polski, które znajdują się w poniższym zbiorze danych:
https://www.kaggle.com/dawidcegielski/house-prices-in-poland?select=Houses.csv

Jakie rozwiązania istnieją?

Podstawowym rozwiązaniem, które będzie konkurencyjne dla zaproponowanego projektu mającego na celu predykcję cen mieszkań będą standardowe wyszukiwarki na popularnych serwisach zajmujących się sprzedażą oraz wynajmem mieszkań. Zaliczyć do nich można między innymi Gumtree czy też Otodom, ale w Internecie istnieją również strony korzystające z modeli, które również mają na celu predykcję cen mieszkań na bazie analizowanych zbiorów danych, do których zaliczyć można między innymi https://obido.pl.

W czym będziemy inni i lepsi?

W Internecie wszelkie projekty służące do predykcji cen są płatne. Nasz projekt będzie pozwalał zrobienie tego za darmo, a ponadto będzie on dosyć dokładny i w oparciu o dostarczone dane pozwoli na predykcję ceny mieszkania w przyszłości bazując na informacjach kumulowanych w bazie. Wyszukiwarki na portalach Gumtree czy Otodom pozwalają jedynie przeglądać obecnie aktualne ogłoszenia, przez co trudno jest dowiedzieć się jak kształtowały się ceny mieszkań w przeszłości, jaki trend w cenach mieszkań występuje.

Korzyści

Jakie korzyści da nasze rozwiązanie?

Możliwość łatwego zorientowania się w sytuacji na rynku mieszkaniowym zarówno dla klienta jak i dewelopera. Dzięki interfejsowi z suwakami i intuicyjnemu przedstawieniu wyjścia modelu jesteśmy w stanie zaproponować łatwą przestrzeń na wyrobienie sobie preferencji co do szukanego mieszkania przedstawiając prognozowane widełki cenowe. Dodatkowo narzędzie można wykorzystać do wyceny konkretnej nieruchomości.

W skrócie - korzyścią jest dokładność predykcji i zaoszczędzony czas użytkownika.

Jak je zmierzyć u klienta?

Czas w którym użytkownik może zorientować się cenowo na rynku zmierzyć można w minutach, które inaczej użytkownik musiałby przeznaczyć na przeszukiwanie portali z nieruchomościami.

Dokładność można zmierzyć za pomocą różnych wartości statystycznych takich jak współczynnik R^2 czy RMSE. Nie są to jednak informacje dla klienta, bardziej opisują to jak poprawne są predykcje modelu, jednak instynktownie klient może poczuć różnicę.

Jaka jest skala korzyści?

Mając na myśl korzyści skali nie jesteśmy obarczeni jakimiś wielkimi kosztami tworząc to narzędzie oraz nie planujemy czerpać z niego zarobku. Zwiększenie skali (liczby użytkowników do których dotarłaby aplikacja) już samo oznaczałoby sukces projektu, którym możemy opisać rozpowszechnienie darmowego narzędzia wśród społeczności.

Realizacja

Cel-czas-nakład (pracy)

Zakładamy, iż na realizację projektu potrzebny będzie około miesiąc. Aby osiągnąć postawiony cel, którym będzie stworzenie modelu służącego do predykcji cen mieszkań dokładnego w przynajmniej 60% przypadków potrzebna będzie regularna praca cotygodniowa. Zakłada się, iż wszyscy członkowie grupy przeznaczą na realizację projektu przynajmniej 2 godziny tygodniowo. Na potrzeby realizacji projektu nie ma wymagania ustalenia nakładów finansowych potrzebnych do realizacji postawionego zadania. Wymagane są jedynie nakłady środków pracy (działanie komputera) oraz pracy żywej.

Jak zmierzyć osiągnięcie celu? KPI

Kluczowym wskaźnikiem, który posłuży do oceny stworzonego modelu będzie współczynnik determinacji R^2, który tak naprawdę pozwala na ocenę dopasowania modelu do wybranych danych. Jeżeli jego wartość przekroczy 0,6, to będziemy mogli uznać, iż cel projektu został wykonany a sam model bardzo dokładnie tłumaczy rzeczywistość.

Jakie cele pośrednie musimy osiągnąć?

Pierwszym celem będzie przygotowanie danych do stworzenia modelu wraz z usunięciem błędnych obserwacji (np. błędnie wpisany rok czy dzielnica). Po ich przygotowaniu możliwe będzie wstępne przedstawienie danych, liczby ogłoszeń w poszczególnych miastach, porównanie rozproszenia cen jak również graficzna analiza ilości ogłoszeń w konkretnie wybranych miejscach i przedstawienie tego w projekcie. Końcowym, głównym celem jest stworzenie modelu służącego do predykcji cen mieszkań i jego analiza.

Jak podzielimy pracę na etapy?

1. Oczyszczenie bazy danych (obserwacje odstające, braki danych, błędy)
2. Wizualizacje
3. Wybór zmiennych do modelu (Współczynniki zmienności, korelacje, istotność)
4. Regresja liniowa-przewidywanie ceny mieszkania na podstawie jego cech i lokalizacji
5. Weryfikacja poprawności modelu
6. Prosty interfejs użytkownika

Zarządzanie ryzykiem?

Jakie są zagrożenia (zewnętrzne)?

Jako, że wykorzystany zestaw danych zawiera ponad 23 tysiące obserwacji i istnieje możliwość łatwego dodania np. najnowszych mieszkań do bazy to nie będzie problemu zbyt niskiej ilości danych.

Może okazać się, że dane nie są niespójne, oraz może występować wysokie korelacje między zmiennymi objaśniającymi i zmienną objaśnianą. Należałoby wtedy pozbyć się potencjalnej współliniowości i uwzględnić w regresji tylko istotne zmienne.

Do modelu powinny być również wykorzystane tylko cechy mieszkań o współczynniku zmienności powyżej 10%. Raczej nie będzie takiego problemu, gdyż przypuszczamy, że zmienne : liczba pokoi, rok budowy, metraż, piętro i wymiary będą dość zróżnicowane.

Przed budową modelu konieczne będzie dogłębna analiza zestawu danych. Będziemy musieli sprawdzić czy w bazie występują braki danych i jeśli tak to zastosować wybraną metodę ich imputacji. W przypadku braków danych na poziomie do 3% obserwacje te zostaną usunięte. Jeśli braków będzie więcej do imputacji wykorzystamy metodę K-najbliższych sąsiadów.

Problemy

W poprzednich projektach wykorzystywaliśmy już potrzebne funkcje i algorytmy.

Problemem może okazać się nierówny podział pracy, ze względu na to, że ciężko jest przewidzieć czasochłonność poszczególnych etapów pracy.

Potencjalnym problemem być również “brak synchronizacji”. Przykładowo do budowy modelu regresji i jego weryfikacji potrzebne będą już oczyszczone dane, a ta część pracy może jeszcze nie być wykonana.
Problemem także będzie zbyt niski współczynnik dopasowania i dalsza analiza problemu okaże się wtedy bezcelowa.

Założenia

Co zakładamy choć jeszcze nie wiemy

Zakładamy że współczynnik R^2 mierzący jak dany model odzwierciedla rzeczywistość będzie wyższy niż 60%.

Jak oceniamy czasochłonność poszczególnych etapów

- Uzupełnienie protokołu (1h)
- czyszczenie danych (2h)
- zrobić coś na wzór UI dla użytkownika końcowego (1-6h)

- lista dzielnic i miast, feature engineering na zmiennej adres tak zeby wyciagnac dzielnice
- analiza danych, statystyki opisowe, wykresy i mapy (3-8h)
- Wnioski końcowe (2-4h)

Jaką ilość czasu chcemy przeznaczyć na projekt - jaka będzie nasza dostępność (tylko szczerze)?

2 godziny tygodniowo każdy z członków zespołu.
Zakładając że mamy miesiąc:
2x4=8 - całkowita ilość czasu jaką poświęci każdy z członków
4x8=32 - całkowita ilość czasu poświęcona na projekt przez wszystkich użytkowników

Decyzje

Jakie decyzje już zapadły i wytyczają kierunek

1. W pierwszej kolejności należało określić zestaw danych na jakim będziemy pracować. Zapadła decyzja o wykorzystaniu danych ze strony https://www.kaggle.com/dawidcegielski/house-prices-in-poland?select=Houses.csv. Są to dane odnoszące się do poszczególnych mieszkań z tych proponowanych na rynku. Jedna obserwacja = jedna oferta

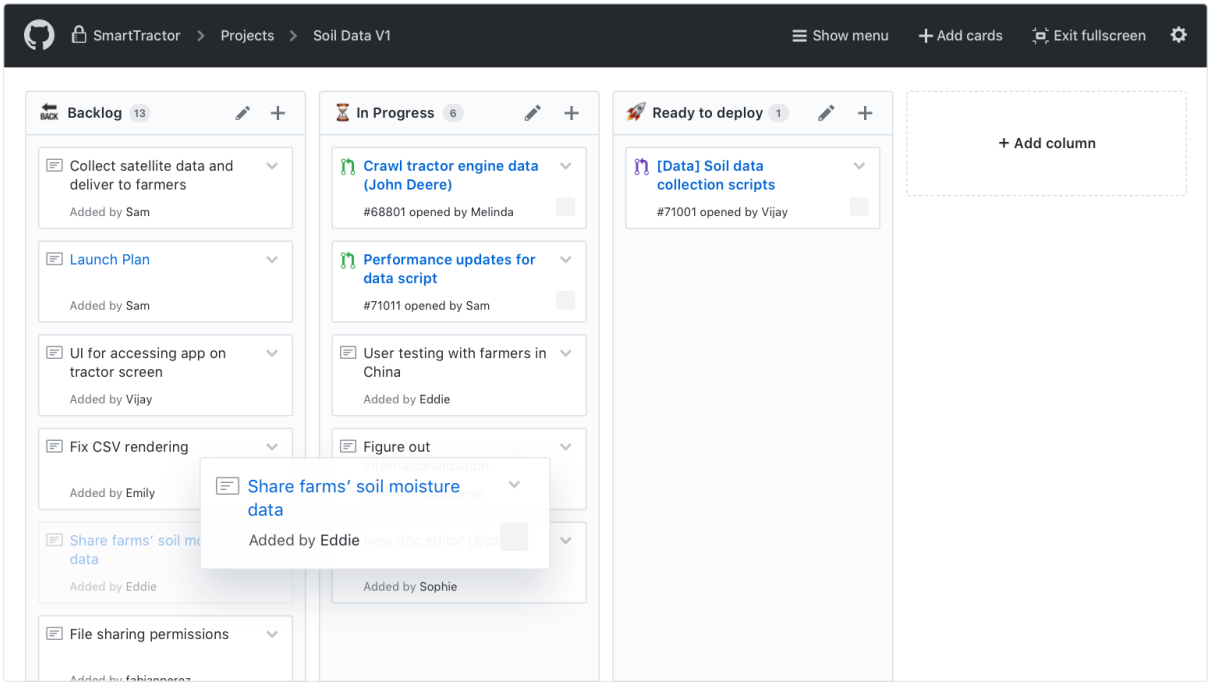
2. Kolejnym krokiem było wyznaczenie celu badań. Zdecydowano się na predykcje cen mieszkań.
3.

Uzupełniać na bieżąco ↑

Działania

Backlog

w sensie coś takiego? (do zrobienia później)



Dalej wg metodologii zarządzania projektem

Notatki

- zrobić coś na wzór UI dla użytkownika końcowego (może być jako suwaczki w jupyterze albo API)
- lista dzielnic i miast, feature enginering na zmiennej adres tak zeby wyciagnac dzielnice
- analiza danych, statystki opisowe, wykresy i mapki
-