НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Методи аналізу геоданих для надання рекомендацій щодо розміщення громадських закладів у Києві

Виконав:
студент групи КА-51
Трохимович М.А.

Керівник:
к.ф.-м.н., доцент
Каніовська І. Ю.

Актуальність дослідження

- Стрімко розвивається напрям “розумне місто” (англ. smart city). Впровадження єдиних електронних квитків на громадський транспорт, соціальні мережі, відцифрування інформації сприяють накопиченню потужної бази знань.
- Аналіз геоданих та побудова математичних моделей, потенційно може покращити доступність товарів та послуг для людей, та збільшити вигоду для бізнесу.

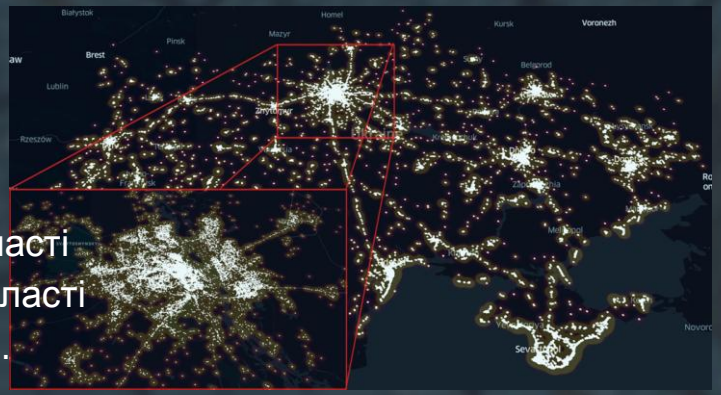
- Об'єкт дослідження
  - Data mining, алгоритми машинного навчання, побудова складних рекомендаційних моделей
- Предмет дослідження
  - Точкові геодані та математичні моделі побудови рекомендацій
- Мета дослідження
  - Дослідити можливість надання рекомендацій, щодо розташування громадських закладів у місті Києві. Побудувати модель для рекомендації геопозиціонування конкретної категорії громадських закладів

Постановка задачі

- Визначити показник успішності громадського закладу
- Проаналізувати дані з різних джерел, спроектувати та видобути необхідні атрибути конкретних локацій
- Побудувати модель для відновлення функціональної залежності успішності громадських закладів від здобутих географічних ознак
- За допомогою моделі знайти цільові значення для нових локацій. Проранжувавши ці значення, дати рекомендацію щодо найкращих місць для розміщення конкретного типу громадських закладів.
- Оцінити якість роботи рекомендаційної системи

Дані та здобуті ознаки

➢ Foursquare

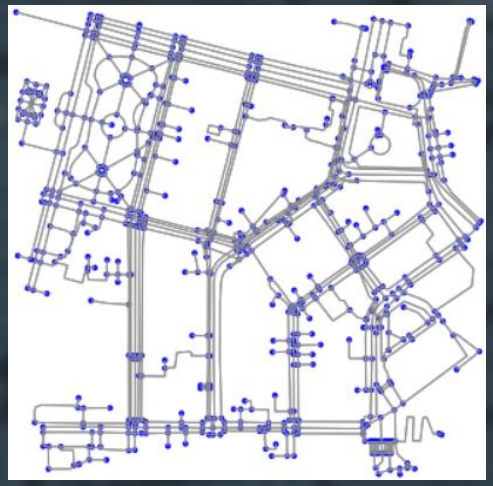


- ❏ Показник успішності – кількість чекінів
- ❏ Density – кількість закладів навколо заданого місця
- ❏ Place entropy - просторова гетерогенність області
- ❏ Competitiveness - конкурентоспроможність області
- ❏ Popularity - загальна кількість чекінів у досліджуваній області
- ❏ Transition Density - мобільність між місцями всередині області
- ❏ Incoming Flow - зовнішній потік людей до заданої області.

➢ EasyWay API

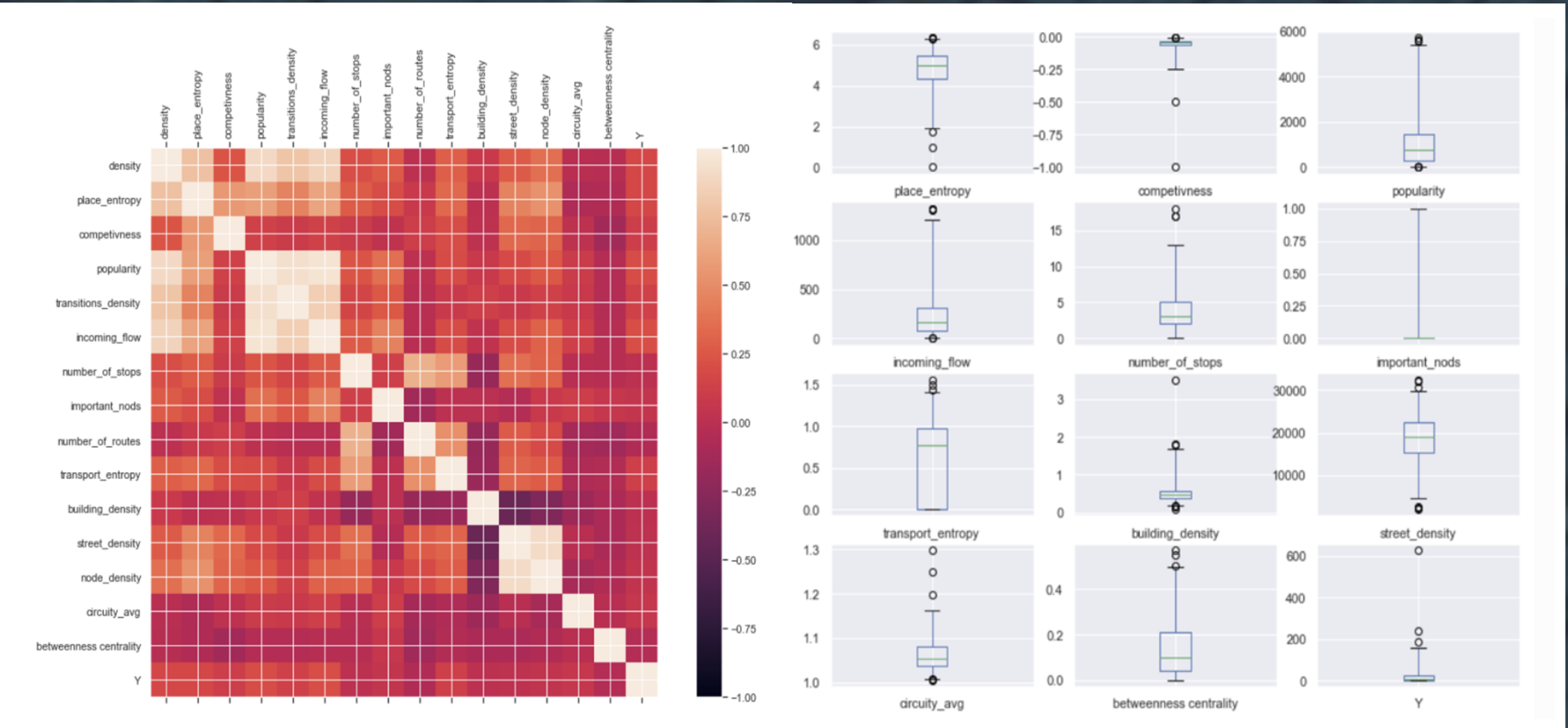
- ❏ stops density – кількість зупинок громадського транспорту
- ❏ important nodes – наявність важливих транспортних вузлів
- ❏ transport density - кількість маршрутів громадського транспорту
- ❏ transport entropy - різноманітність видів громадського транспорту

➢ OSMnx API

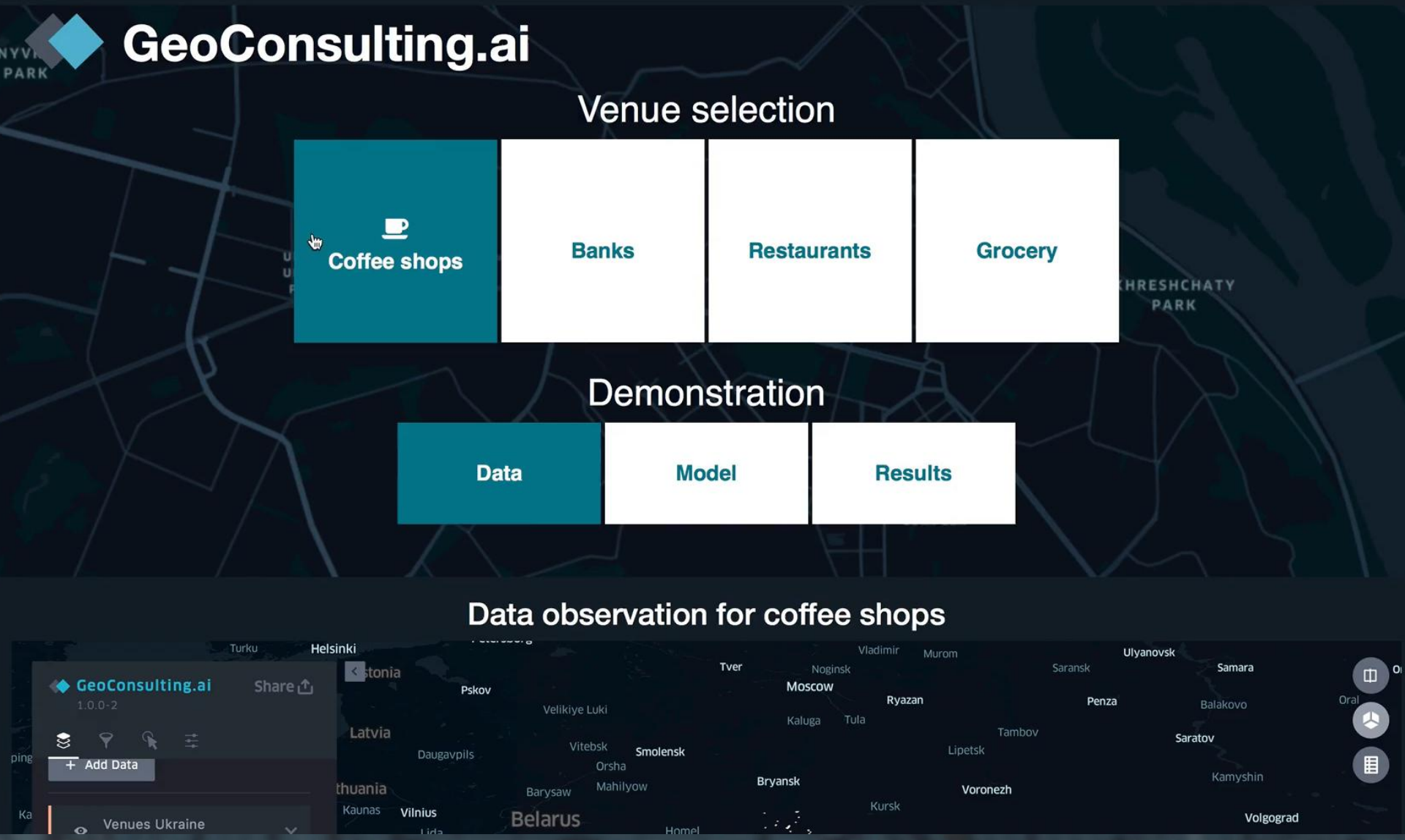


- ❏ street density – щільність вулиць
- ❏ intersection density – щільність перехресть
- ❏ average circuity - середня заокругленість вулиць
- ❏ betweenness centrality - важливість даної вершини у мережі

Огляд здобутих ознак



Демонстрація в графічному інтерфейсі



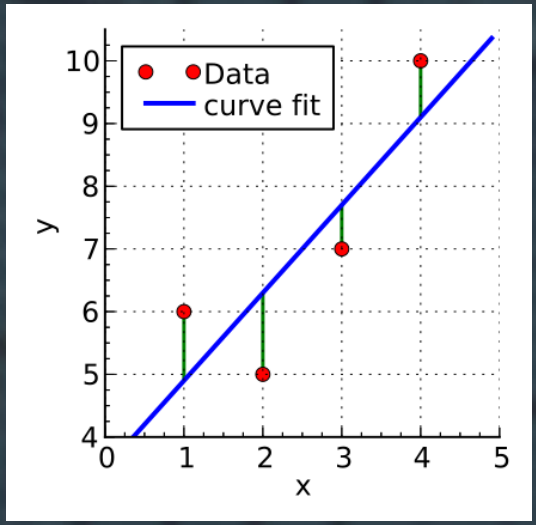
Відновлення функціональної залежності

❑ Лінійні моделі

Лінійна регресія:

- Стохастичний градієнтний спуск
- Ridge (L2 регуляризація)
- Lasso (L1 регуляризація)
- ElasticNet (L2+L1 регуляризація)

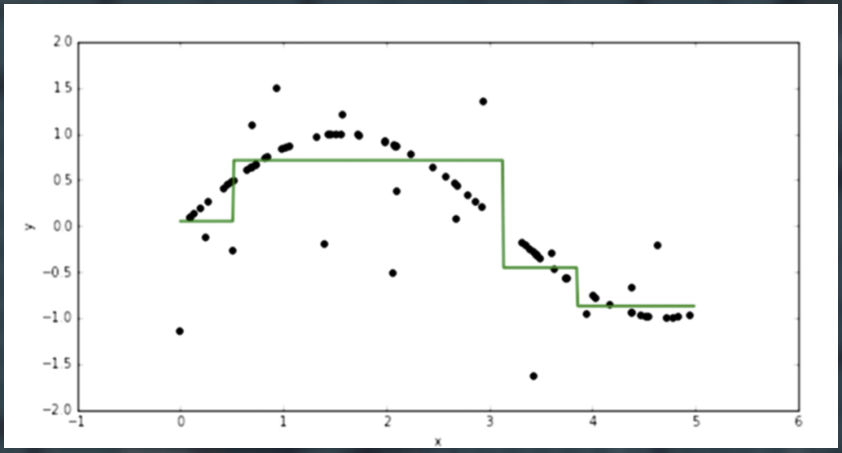
$$Q(w, X) = \frac{1}{l}||Xw - y||^2 + \lambda\,||w||_{l_1} + \lambda||w||^2 \rightarrow min_{w \in R^d}$$



★ Нелінійні моделі

Decision tree based algorithms:

- Decision Tree
- Random Forest
- XGBoost Regression



Decision Tree and Random Forest

- Навчання моделі:
  - Існує вершина m у якій Xm об'єктів з навчальної вибірки.
  - Відбувається її поділ на $X_l = \{x \in X_m | \, [x^j \leq t]\}, X_r = \{x \in X_m | \, [x^j > t]\}$

    При цьому параметри $[x^j \leq t]$ обираються з таких міркувань:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} H(X_l) + \frac{|X_r|}{|X_m|} H(X_r) \to \min_{j,t}$$

$$H(X) = \frac{1}{|X|} \sum_{i \in X} (y_i - \overline{y}(X))^2, \text{ ae } \overline{y} = \frac{1}{|X|} \sum_{i \in X} y_i$$

- Можливі критерії зупинки:
  - У вершині тільки один об'єкт навчальної вибірки
  - Глибина дерева досягла певного значення
  - Кінцева відповідь моделі в листі m: $a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$
- Random Forest це сукупність дерев, відповіді яких комбінуються.

Метрики

- Precision at K

$$p@K = \frac{\sum_{k=1}^{K} r^{true}\left(\pi^{-1}(k)\right)}{K} = \frac{\text{number of relevant}}{K}$$

- Average precision at K

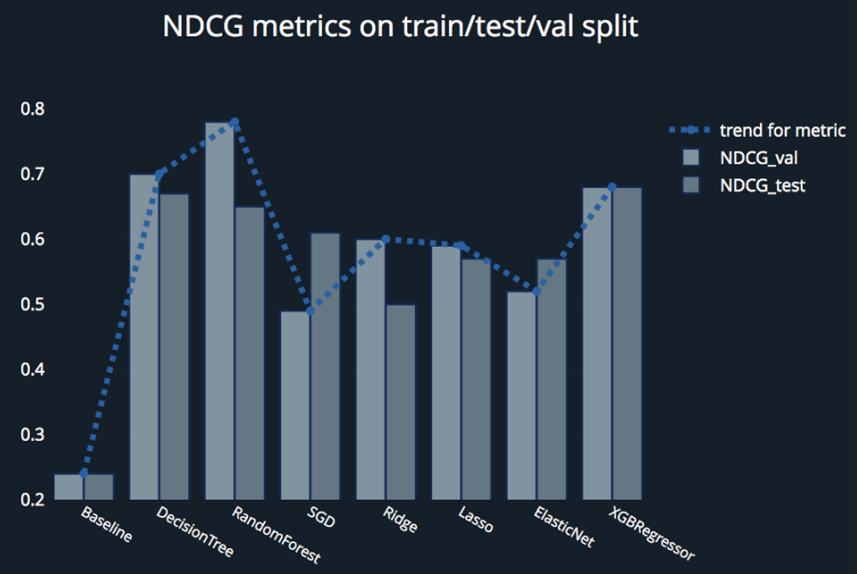
$$ap@K = \frac{1}{K}\sum_{k=1}^{K} r^{true}\left(\pi^{-1}(k)\right) \cdot p@k$$

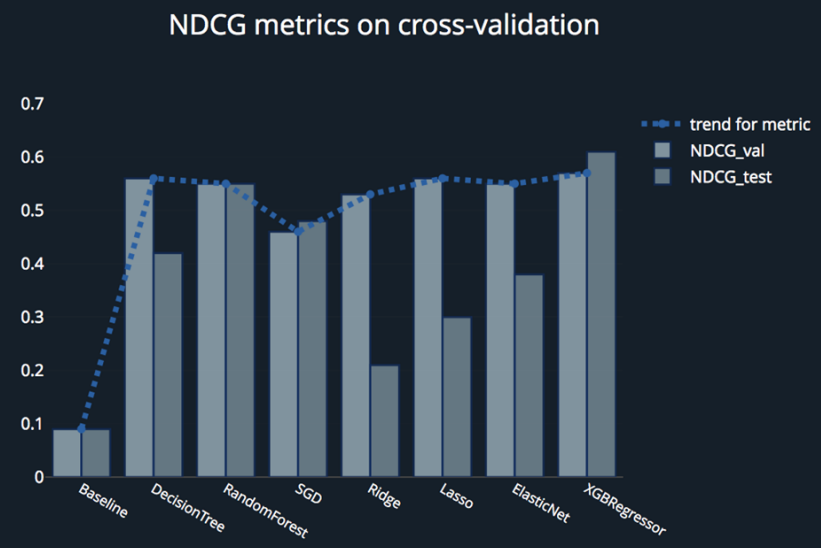
- Normalized discounted cumulative gain (nDCG)

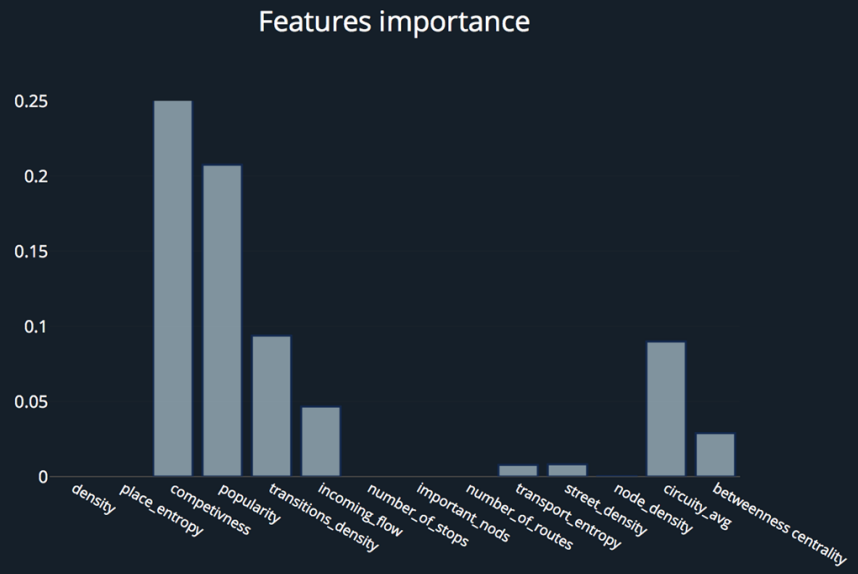
$$DCG@K = \sum_{k=1}^{K} \frac{2^{r^{true}\left(\pi^{-1}(k)\right)} - 1}{\log_2(k+1)}$$

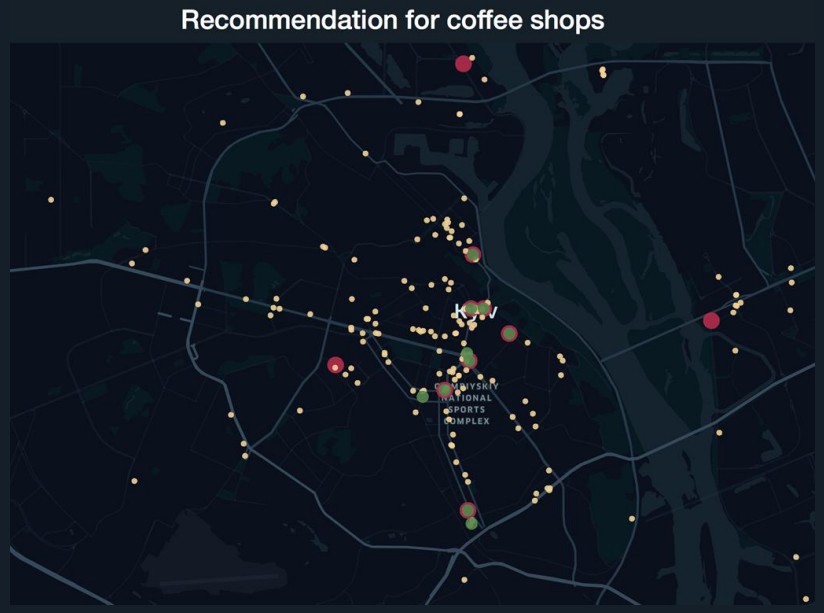
$$nDCG@K = \frac{DCG@K}{IDCG@K}$$

Результати експериментів з кав’ярнями









Демонстрація роботи моделей

GeoConsulting.ai

Venue selection

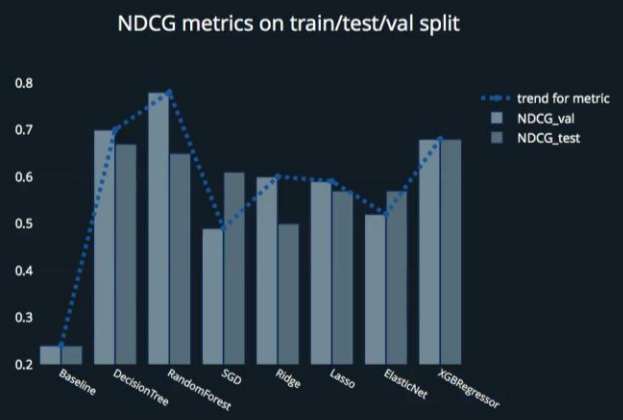
Coffee shops | Banks | Restaurants | Grocery

Demonstration

Data | Model | Results

Models comparison

	NDCG_val	NDCG_test	AP_val	AP_test	Precision_val	Precision_test	diff_in_%
eline	0.24	0.24	0.09	0.09	0.3	0.3	0.0
isionTree	0.7	0.67	0.52	0.48	0.6	0.6	3.48
domForest	0.78	0.65	0.65	0.47	0.7	0.6	16.49
D	0.49	0.61	0.29	0.41	0.4	0.5	22.58
ge	0.6	0.5	0.39	0.29	0.5	0.4	16.42
so	0.59	0.57	0.39	0.35	0.5	0.5	3.87
sticNet	0.52	0.57	0.32	0.34	0.4	0.5	8.73
BRegressor	0.68	0.68	0.5	0.48	0.6	0.6	0.99



Практичне застосування

- Рекомендація щодо розміщення
  - Визначення найкращого географічного положення нового громадського закладу
- Оцінка ефективності роботи бізнесу
  - Знаючи характеристики конкретної локації можна визначити теоретичний показник її успішності. Порівнявши його з емпіричним показником можна оцінити ефективність організації роботи громадського закладу

Висновки

- Проаналізували дані з різних джерел, спроектували та видобули необхідні атрибути конкретних локацій, знайшли закономірності в даних.
- Побудували моделі для відновлення функціональної залежності успішності громадських закладів від здобутих географічних ознак.
- За допомогою моделі знайшли цільові значення для нових локацій. Проранжувавши ці значення, дали рекомендацію щодо найкращих розташувань кав’ярень у місті Києві.
- Спроектували рекомендаційну систему та програму для візуалізації, демонстрації та інтерпретації результатів дослідження.

Подальші дослідження

- Використання нейронних мереж для відновлення функціональної залежності.
- Використання більшої кількості джерел даних.
- Порівняння ефективності роботи у різних містах та на різних типах закладів.

Дякую за увагу!