GEO. Module

Если мы что-то затеваем (назовем для простоты стартапом), то ключевые компетенции должны иметь внутри команды, причем на уровне близком к экспертному.

Наши компетенции:

- умеем строить модели (другие тоже умеют)
- понимаем источники внешних данных (тут, опять же, мы не одиноки)
- умеем внедрять модели, сервисы (хорошо, но другие тоже умеют)
- понимаем бизнес-логику (уже лучше)
  - страхование (на “отлично”)
  - можем разобраться в других не узкоспециализированных сферах
  - понимаем отдельные процессы (например: dwh)
- понимаем экономику и финансы (совмещение с ml - хорошо)
- понимаем стратегический менеджмент (круто, но применить бы)

Input data

For the task of geographic segmentation, need to determine sets of addresses:

1. Customers' registrations addresses
2. Sellers' addresses
3. Loss points' addresses
4. Other possible geotagging trainers in the process of the client's life

Further, the words cluster, segment, group should be understood as the same thing. The task of geographic clusters should be considered from the point of view of both geo-risk and the potential for growth in sales, in particular, vacant unoccupied territories with a high population

density with different wealth.

In this project, you need to work with data for all types in the following sequence:
• compulsory - third party liability, fully comprehensive insurance
• in case of successful solution of the problem of auto insurance: Property, health accident, travelers insurance, medicince, corporate block

Insurance company uses many addresses in various business processes. In particular, at the stage of insurance for various types, clients fill out the registration addresses of owners, policyholders. The addresses of losses are recorded when claiming losses for various types of insurance.

When insuring property, the addresses of the objects are filled in. It is an important source of information when properly processed.

Also in this problem it is required to consider the applicability of the geo2vec algorithms (loc2vec, place2vec ...)
for questions of defining geographic clusters. It is required to develop a territory coding algorithm possessing geographical knowledge of neighboring territories.

The resulting vector representations (geoembedings) of territories should be further used for tariffication models

What problem are we solving and why?
The goal of task 2 of the project is to identify geographic clusters. Show the applicability of this geographic segmentation

What needs to be done?

1. Show borders of clusters on map
2. Together with the company's employees, select data sources with geographic information.
3. Choose the optimal technology stack for the company to solve this problem. Agree with the customer. If necessary, you can use open source GIS software for Creation of a GIS Database
4. Using the selected stack, prepare the data for the study.
5. If necessary, enrich the addresses with geocoordinates using the available services. Codes of Russian classifiers.
6. Conduct geographic clustering using various methods - density, centroids, distributions and others.
7. Explain the chosen approaches to the customer
8. Show the applicability of the study to improve unprofitability by building work with selected segments, for example, by identifying geofences that are significant for predicting losses.
9. Show the applicability of clustering to increase the company's sales, for example, by searching for free zones where the company is poorly represented.
10. Check clusters for homogeneity with a high effect on forecasting models.
11. Show the potential of using clusters to build scoring, tariff rates and search for fraud.
12. Analyze the change in geoclusters in dynamics
13. Compare the resulting clusters with similar solutions of insurance companies on the market
14. Provide convenient access for analysts to view geoclusters
15. Carry out prototyping and building models and services.
16. Create a working prototype of the geographic cluster definition service.

What is the result of the work?
What criteria will be used to evaluate the result?
What help can you count on?
What skills are needed?
Skills of working with GIS systems

What will the load on the project be?
What will be the format of interaction?

What common sources of knowledge can be viewed on this task?

https://www.mdpi.com/2227-9091/7/2/42/htm

https://www.esri.com/~/media/Files/Pdfs/library/fliers/pdfs/location-analytics-insurance.pdf

https://www.mdpi.com/2220-9964/8/3/134

https://arxiv.org/pdf/2005.01690.pdf

https://geog.ucsb.edu/~jano/place2vec.pdf

https://www.sentiance.com/2018/05/03/venue-mapping/

https://journals.udsm.ac.tz/index.php/orsea/article/viewFile/827/769

Other ideas to be done:

- Ask Autostat about details zip code statistics
- Connection to agent geography
- Comparison betwenn car price and flat price
- General road conditions, topography, signage, and

  other distractions can have a profound effect on occurrence of accidents

  hence the resulting claims.
- Commute time, distance to shopping centers,

  availability of public transit, and housing affordability are some of the factors

  that determine the choice of where to live. Some of those same factors

  also have a big impact on where, how often, and how far an insured drives.

  Annual mileage is a strong predictor of risk for auto insurance and is

  animportant rating variable in many rating plans
- crime data, police stations
- Calims types, europrotocol analysis

Business processes uses geographical information

Business process name and case description	**Input data**	**Impact**	**Out data**	**Quality estimation**
Pricing for territories	Owner address Insurer address	Calculate cluster for territory and its properties, estimate of vehicle usage area	Put cluster as factor to model	Increasing quality of glm models, basically deviance And Gradient boosting risk models
General territory risk assesment, distance factors engineering	Region, city, country, town	Estimate risk by using customers territory properties relative to major city, roads and	Risk estimation	Reduce deviance in risk prediction

		other objects on map, it also can be weighted average between territories. It is important to split systematic risk and non systematic risks. And estimate deviance relative other features to estimate dependencies.		
Reserves by territory	Claims dynamics in history	More accurate calculation of reserves	Reserve prediction	Metrics for reserve prediction
Claims map risk assesment	Claim address	Estimate of claims density for terriotry	Estimate risk by claims location	Context embedings comparison Homogenity Davies-Boulding index Silhouette Coefficient Other clustering metrics
Repair station settlement	Customer data Repair stations locations and properties	Clients traffic distribution	Repair station recomendations	Cost decreasing
Geo marketing	Owner address Insurer address Telecom operators data Context advertisment	Define probability of product buying or office visitig, routing. Estimate penetration on territory for company and understand possibilities to catch market Estimate transition of office properties because of online channel development	Probability of product buying	Get new clients by territory Attractions cost decreasing
Geo product development	Customer data Geo data Channel data	More probable products and options	List of offers by territory	Increase conversion Increase average check (serveral premiums) LR decreasing
Telematics	Device or phone data Customer data	Telemetics services, Product options/discounts	Scores for differnet tasks	Increase conversion Increase average check (serveral premiums) LR decreasing
Phone call tracking geo (if possible)	Telecom operator data	Put call to geo area	Risk definition by different adresses of calling and	?

			registered data	
Sales agents probability	Owner adresses Insurer addresses	Find better areas for agents	Clusters to work with	Increase sales in that clusters
Sales channels transformation	Sales points descriptions Cost of business	Change structure of sales by area	Sales points closing or opening recomendation Sales channels transformations	Cost of sales
Usage data from GIBDD	GIBDD data Owner adresses Insurer addresses	Calcualate density of GIBDD data	Density of claims by area	Metrics of risk prediction models
Facotors for process models	Penetreation factors	Impact on probability of • subrogation • agreement • appeal • court	Probability of process models	Increase metrics of supervised models
Sales process for new customers	Quotes data Geo data Sales person data	Different products and options for different geo areas	Recommended product options by geo areas	Increase conversion Increase average check (serveral premiums) LR decreasing
Claims settlement properties by area	Cost of claims by area and repair stations	Change routes for accidents settlement	Recomendation for spare parts cost	Reduce cost of repairs
Working with local gibdd service for accident decreasing	Accident data	Set up road lights or something	Recomendations	Accidents decreasing

Using geoservices to show maps and clusters

.....

I Indicators of penetration into the territory.

Sales.

- Penetration of the product by the number of the vehicle fleet (is there any detailing finer than the region?).
- Penetration of the product (cross-selling) from the population.
- The number of agents (intermediaries) from the population. Average number of customers (residents) per agent (intermediary).
- Average distance between agents (intermediaries).

Call center.

- Number of hits from sales volume.
- Time of calls.
- Call duration.
- Subject of the appeal.

- In general, all indicators of CC by territories.

Marketing.

- Elasticity of demand.
- Influence of advertising campaigns.

Settlement.

- Density of workshops (delir / non-dealer).
- The density of the settlement company employees.
- The number of staff to settle on the volume of the portfolio.
- The number of settlement staff on the size of the park.
- Number of workshops per policy
- Number of ships per unit area or population in relation to territory
- Malice of courts and judges on the territory

II Impact on unprofitableness. Territory data.

- Average values for the territory within the radius.
- The number of inhabitants in the radius.
- The status of the settlement.
- Distance to the nearest regional center.
- Distance to your regional center
- Distance to the nearest regional center.
- Distance to your regional center.
- Which is closer: the district or regional center
- Population density in the radius.

Impact on unprofitability (data on the company's presence in the territory).

- Distance from the place of sale to the place of residence.
- Are there agents closer.

III Portrait of the territory.

- Number of settlements.
- Aveage size of the settlement.
- Average distance between settlements.
- Population density in the radius.
- Distance to the nearest regional center.
- Distance to your regional center.
- Distance to the nearest regional center.
- Distance to your regional center.
- Which is closer: the district or regional center.
- Density of the agent network. o Density of the dealer network.

- The size of the vehicle fleet per capita.
- Density of auto lawyers.
- Density of alcoholic beverages.
- Density of entertainment centers.
- Density of shopping centers.
- State of the road network.
- Development of public transport.
- Development of cycling.
- Density of small businesses.

IV Telecom data.

- Traffic profile past the point of sale.
- Behavior of residents of the territory (frequency and distance of travel,…).\

V Antifraud.

- Average density of fraudsters.
- The state of the judicial system.
- Number of court proceedings on corruption articles.
-

For task of metrics quality estimation we understand that it isn ot easy to estimate metricvs for unsupervised learing. Because data may be not have convexity. . If your metric assumes convexity but the data is naturally non-convex, then the metric is useless for that algorithm.

Additional files

Boskov and Ver…....pdf
04 сент. 2020, 05:51 AM

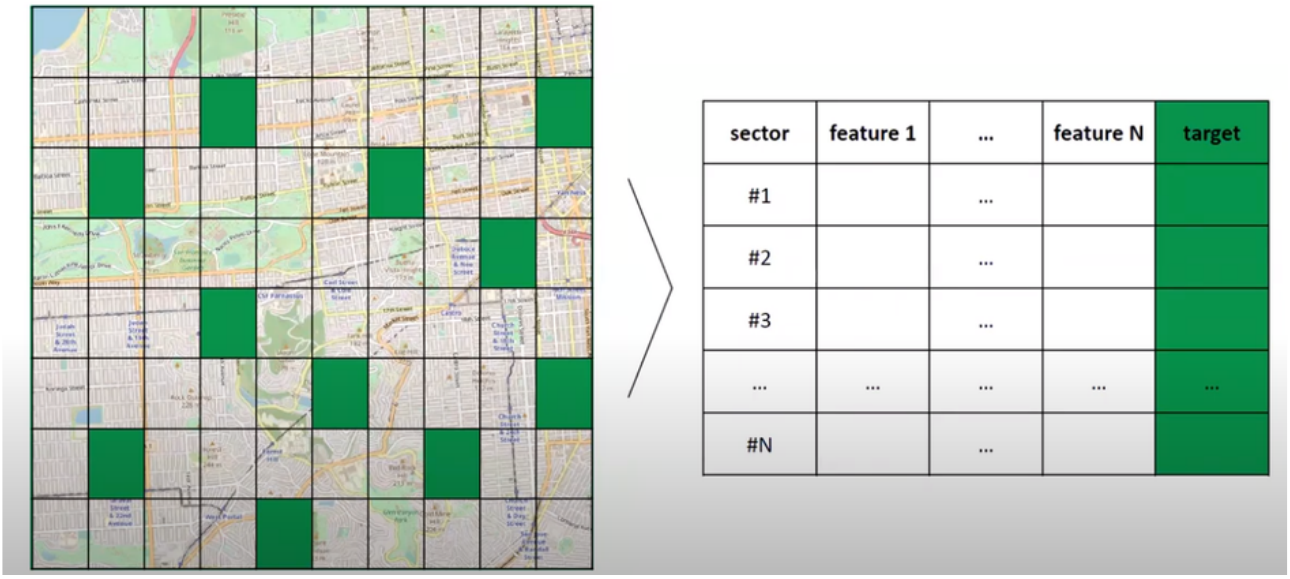
Anderson - Det… ….pdf
04 сент. 2020, 05:51 AM

reorpaфияglm.pdf
26 авг. 2020, 09:17 PM

---

https://www.youtube.com/watch?v=de81Ev-97aI&feature=emb_title

В Сбербанке используется подход генерации фичей по территории / сектору

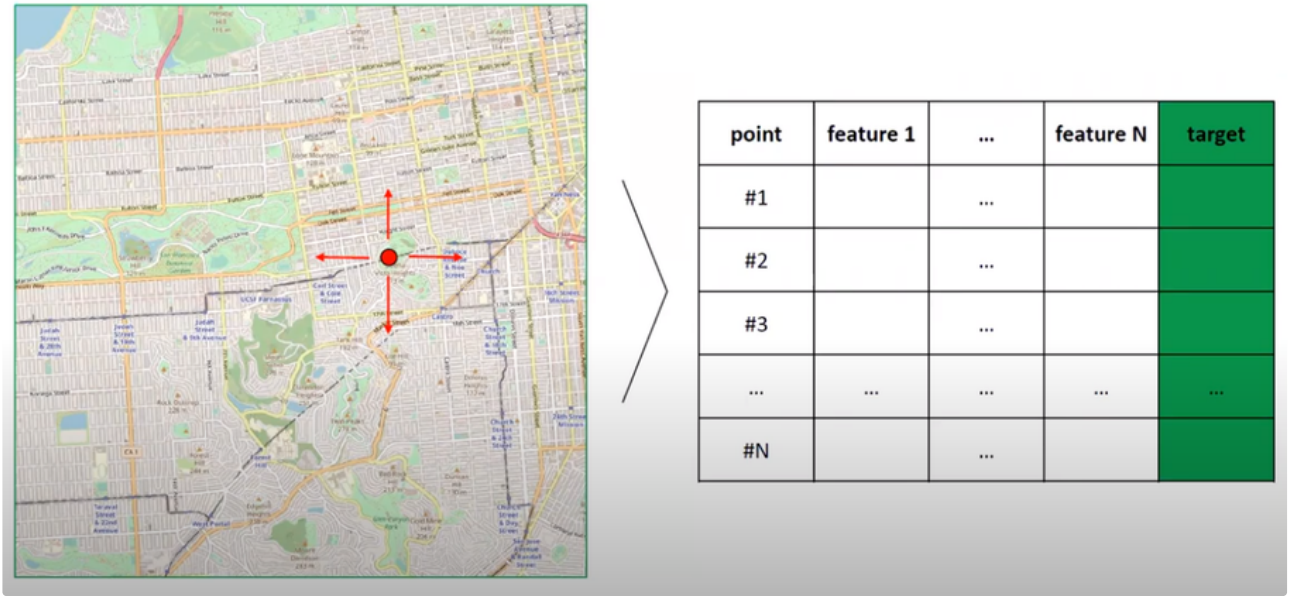


sector	feature 1	...	feature N	target
#1		...		
#2		...		
#3		...		
...	...	...	...	...
#N		...		

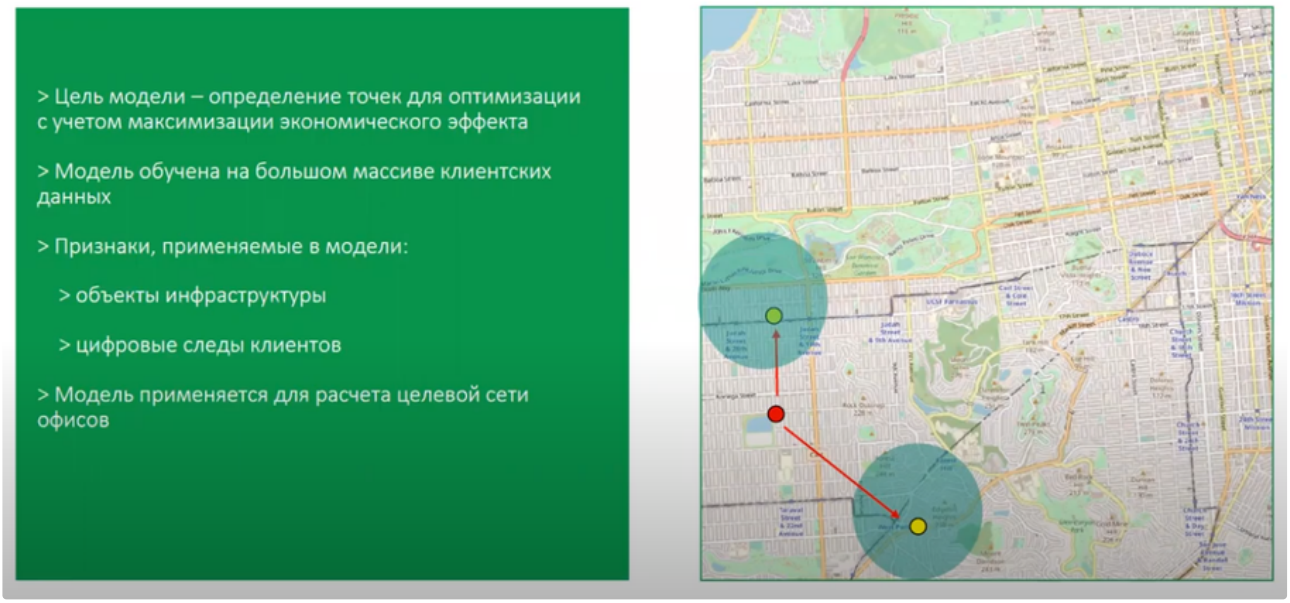
Также используются в задаче перемещения офисов признаки основанные на основе расстояния до объектов инфраструктуры и прочих объектов для

Так как при перемещении рассматривается изменение клиенто-потока в новой точке

Обучающая выборка строится на основе существующих точек расположения офисов и определения ключевых признаков влияющих на показатели подразделения



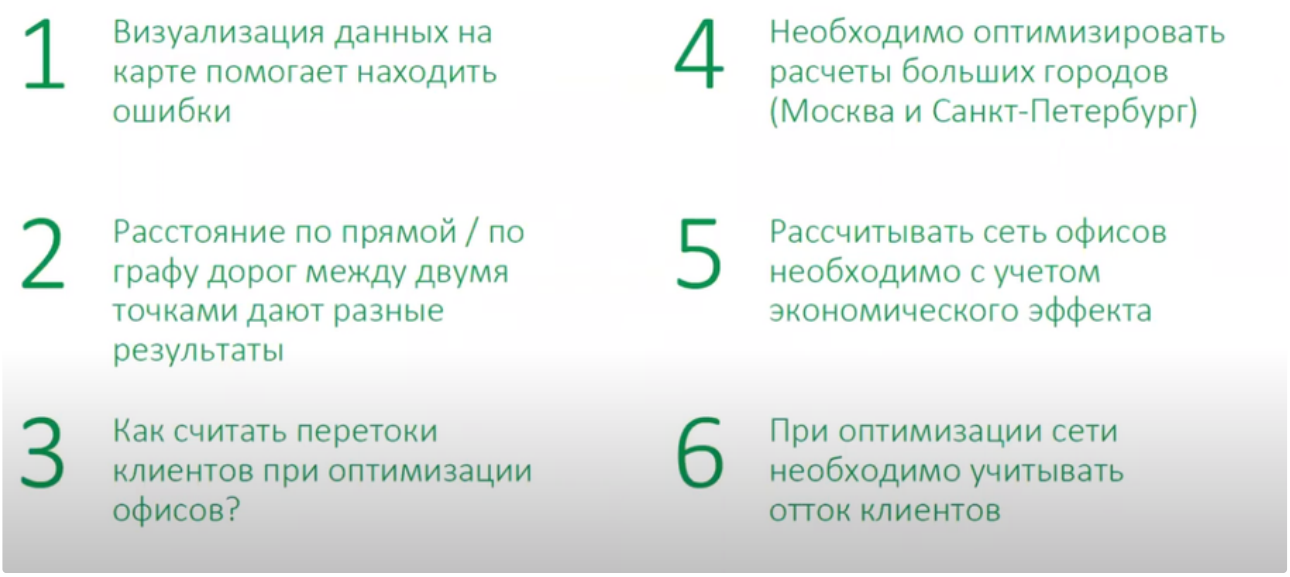
point	feature 1	...	feature N	target
#1		...		
#2		...		
#3		...		
...	...	...	...	...
#N		...		

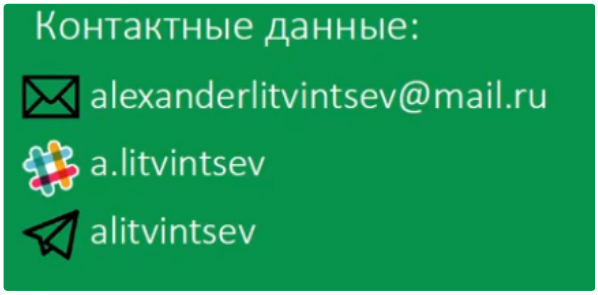


Определение целевой сети офисов

Большой объем данных можно обойти путем семплирования

Важно учитывать природные ограничители





Для обсуждения подобной задачи в Сбербанке.

1. Модель для открытия подразделений. Агрегирование признаков по транзакциям на основе секторов, часть локаций, где есть офисы образует обучающую выборки. Есть сектор, признаки сектора, таргет показатели подразделения на территории. Затем по модели прогнозируем по другой территории.
2. Модель для перемещения подразделений. Строится модель по признакам расстояниям до ближайшей инфраструктуры и прочим окружающим точкам. То есть нет разбиения на сектора, но есть точка - открытый офис и показатели этой точки в виде таргетов, а признаки окружения точки это фичи.
3. Модель максимизации экономического эффекта точки продаж путем перебора локаций.

Geo features for insurance

Level	General external data	Local external data	Internal insurance data
Macroregion	population density		average frequency (claims, courts, …) trend seasonal coefficients
Region	population population density ➕ roads length and density http://fedstat.ru/indicator/56281 CBR stat (http://cbr.ru/insurance/reporting_stat/ http://cbr.ru/insurance/analytics/) region accidents places https://fedstat.ru/indicator/60241 https://fedstat.ru/indicator/59311 deaths on roads https://fedstat.ru/indicator/59114 GRP per capita Gibdd accidents	vehicle fleet (amount & structure) oficial statistics (unemployment, average salary, …)	average frequency (claims, courts, …)
City / Town / District	population population in the circle ➕ distance to the regional center ➕ distance to the nearest big city ➕ population of the nearest big city ➕ road quality by photo	presence in the circle (highways, rivers, bridges, railways, …) oficial statistics (accidents, elections, crime, demography, unemployment, average salary, …)	average frequency (claims, courts, …) in the circle number of policies in the circle courts, lawyers

	https://habr.com/ru/post/437542/		
Local	loc2vec	density of accidents WalkScore pedestrian traffic length to metro from home quality of public transport	
Clients		telecom data credit cards transactions average receipt amount locations & movements (telecom, gps)	