

Развитие инструментов предиктивной аналитики в целях повышения эффективности мониторинга проектов в сфере жилищного строительства

Ефремов Сергей

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра технологий цифровой трансформации

Научный руководитель канд. экон. наук, доц. А;А. Помулев

Москва,
2022 г.

Проблема

Для определения потенциально возможных финансовых рисков необходим мониторинг проектов в сфере жилищного строительства.

Оценка даты фактической готовности

Предлагается строить прогноз ключевого макропараметра объекта жилищного строительства — даты фактической готовности.

Предложение

Проанализировать признаковое описание объектов строительства и построить эффективную регрессионную модель, способную предсказывать дату фактической готовности.

Работы по проектному финансированию

- *Kipf T. N., Welling M.* Semi-Supervised Classification with Graph Convolutional Networks // Proceedings of the 5th International Conference on Learning Representations, 2017
- *Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P. S.* A Comprehensive Survey on Graph Neural Networks // IEEE Transactions on Neural Networks and Learning Systems, 2020

Работы по алгоритмам машинного обучения

- *Derevyanko G., Grudinin S., Bengio Y., Lamoureaux G.* Deep convolutional networks for quality assessment of protein folds // Bioinformatics (Oxford, England), 2018
- *Pagès G., Charmettant B., Grudinin S.* Protein model quality assessment using 3D oriented convolutional neural networks // Bioinformatics (Oxford, England), 2019
- *Baldassarre F., Menéndez Hurtado D., Elofsson A., Azizpour H.* GraphQA: Protein model quality assessment using graph convolutional network // Submitted to Bioinformatics, 2020

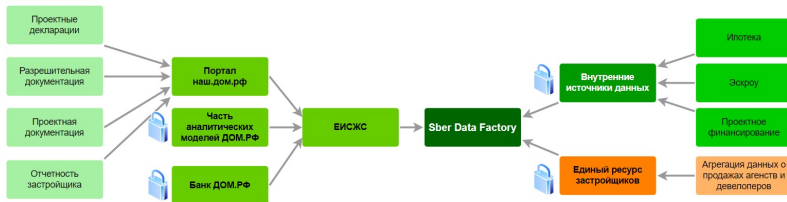


Схема источников данных

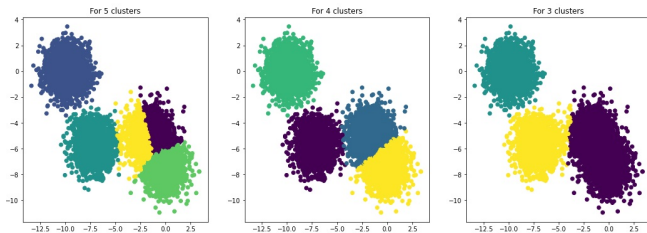


Схематичное изображение исследованных моделей

Метод k -средних

$a(x_i) = \operatorname{argmin}_{1 \leq k \leq K} \rho(x_i, c_k)$ — критерий принадлежности к кластеру;

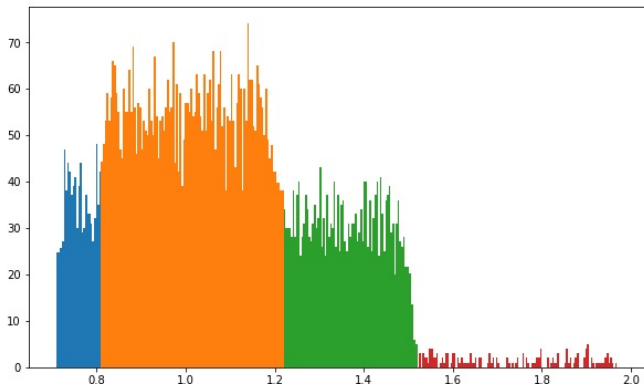
$$c_k = \frac{1}{\sum_{i=1}^I [a(x_i) = k]} \sum_{i=1}^I [a(x_i) = k] x_i \quad \text{— центр кластера}$$



Результат работы k -means для разного количества кластеров

Пороговая классификация

$$a_i \leq \frac{p_j \cdot |R|}{\sum_{k \in R} p_k} \leq a_{i+1}; \quad a_1 \approx 0,8; a_2 \approx 1,2; a_3 \approx 1,5.$$



Распределение значений признака f

Модель	Внутреннее расстояние	Внешнее расстояние	Индекс Данна
$k - means^{++}$ 5	18,7	28,9	1,15
$k - means^{++}$ 4	15,9	28,1	1,31
$k - means^{++}$ 3	19,4	23,9	1,05
Пороговая классификация	20,2	27,4	1,23

Сравнение метрик качества $k - means^{++}$ для разного количества кластеров

Линейная регрессия

$$\frac{1}{l} \sum_{i=1}^l (\langle \omega, x_i \rangle - y_i)^2 \rightarrow \min_{\omega}.$$

Градиентный бустинг

$$a_N(x) = \sum_{n=0}^N \gamma_n b_n(x) - \text{взвешенная сумма базовых алгоритмов}$$

$$\sum_{i=1}^l L(y_i, a_{N-1}(x_i) + \gamma_N b_N(x_i)) \rightarrow \min_{b_N, \gamma_N}.$$

MSE

$$MSE(a, X) = \frac{1}{I} \sum_{i=1}^I (a(x_i) - y_i)^2$$

R^2

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^I (a(x_i) - y_i)^2}{\sum_{i=1}^I (y_i - \bar{y})^2}$$

MAE

$$MAE(a, X) = \frac{1}{I} \sum_{i=1}^I |a(x_i) - y_i|$$

MAPE

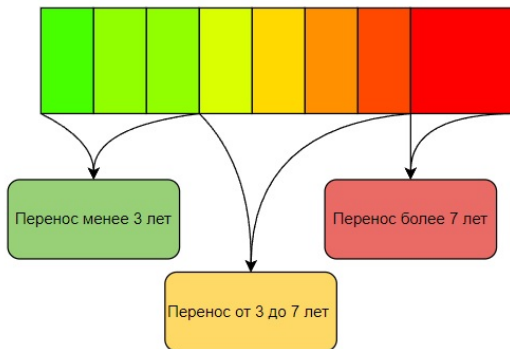
$$MAPE(a, X) = 100\% \times \frac{1}{I} \sum_{i=1}^I \frac{|a(x_i) - y_i|}{|y_i|}$$

Модель	MSE	R^2	MAE	MAPE
LinearRegression	61564	0,516	307	20%
$k - means^{++}$ & LinearRegression	45671	0,634	260	17,7%
f-clas & LinearRegression	43282	0,652	248	17,3%
f-clas & LightGBM	24467	0,808	224	13,6%
LightGBM	22009	0,827	196	12%

Показатели качества построенных моделей

Мониторинг портфеля объектов строительства

Результат работы модели на первое июня показал, что доля объектов в красной зоне около 1,5%, а объектов в желтой зоне порядка 20%



Градиентная визуализация

Необходимость мониторинга

- Риски только по одному объекту могут составлять более 25 млрд. руб.
- Ручной мониторинг отделом из 10 специалистов обойдется порядка 30 млн. руб ежегодно.
- Отношение составляет менее промилле от потенциальной суммы риска только по одному проекту.

Плюсы автоматизации

- Возможность сэкономить около 80% трудового ресурса аналитиков
- Стоимость внедрения решения не превышает 5 млн. руб. при существовании возможности масштабирования.
- Стоимость обслуживания около 1 млн. ежегодно с учетом всего жизненного цикла.

Полученные результаты

- Разработаны методы прогнозирования даты фактической готовности объектов жилищного строительства
- Предложены методы предварительной предобработки данных для повышения качества прогноза модели
- Результаты работы алгоритмов показывают состоятельность подхода и репрезентативность результата

Дальнейшие исследования

- Возможно обобщение модели на другие важные для проектного мониторинга признаки
- Переход к прогнозированию динамически изменяющихся показателей
- Обогащение признакового пространства показателями импортозависимости