

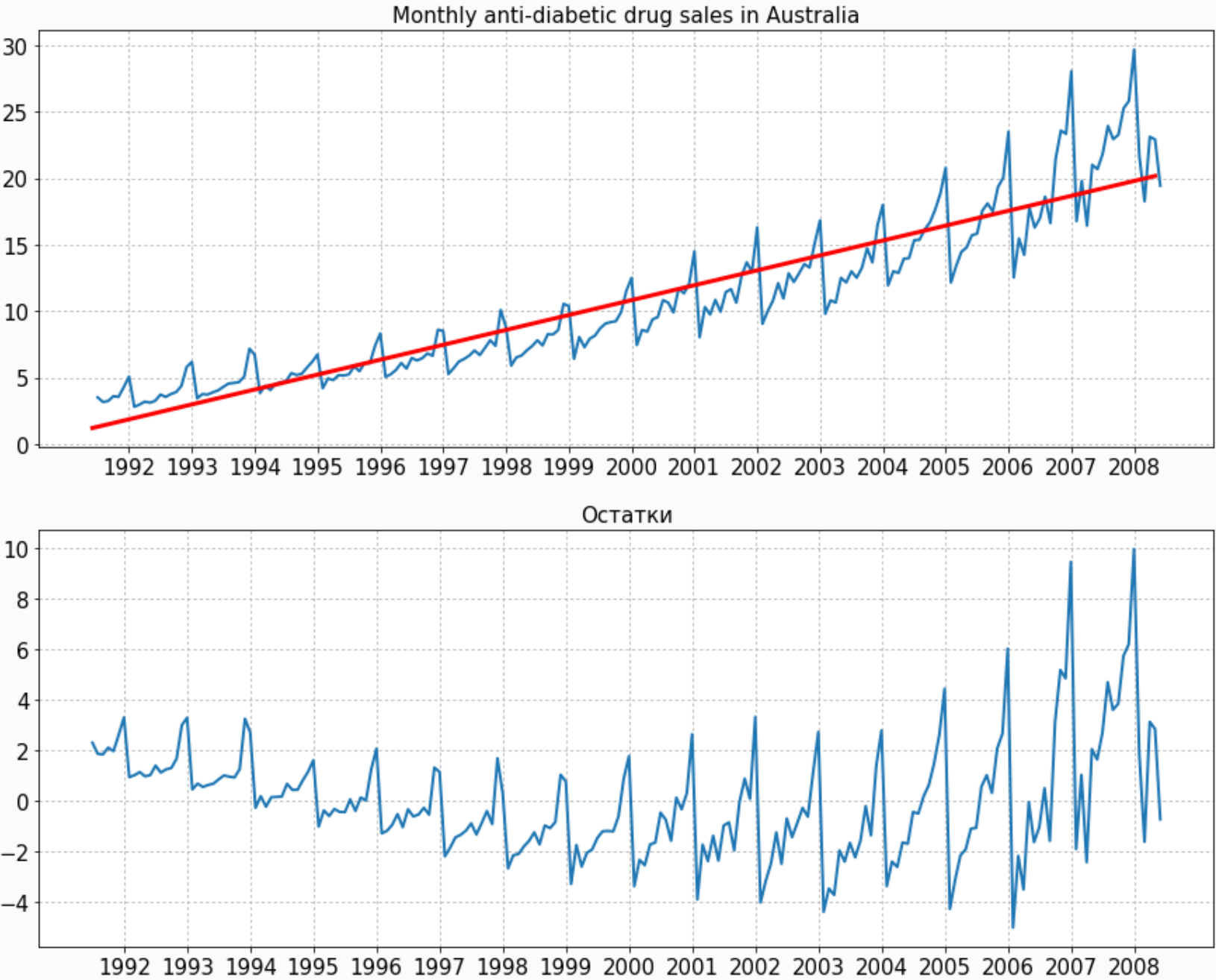
Применение классических

методов для решения задач прогнозирования

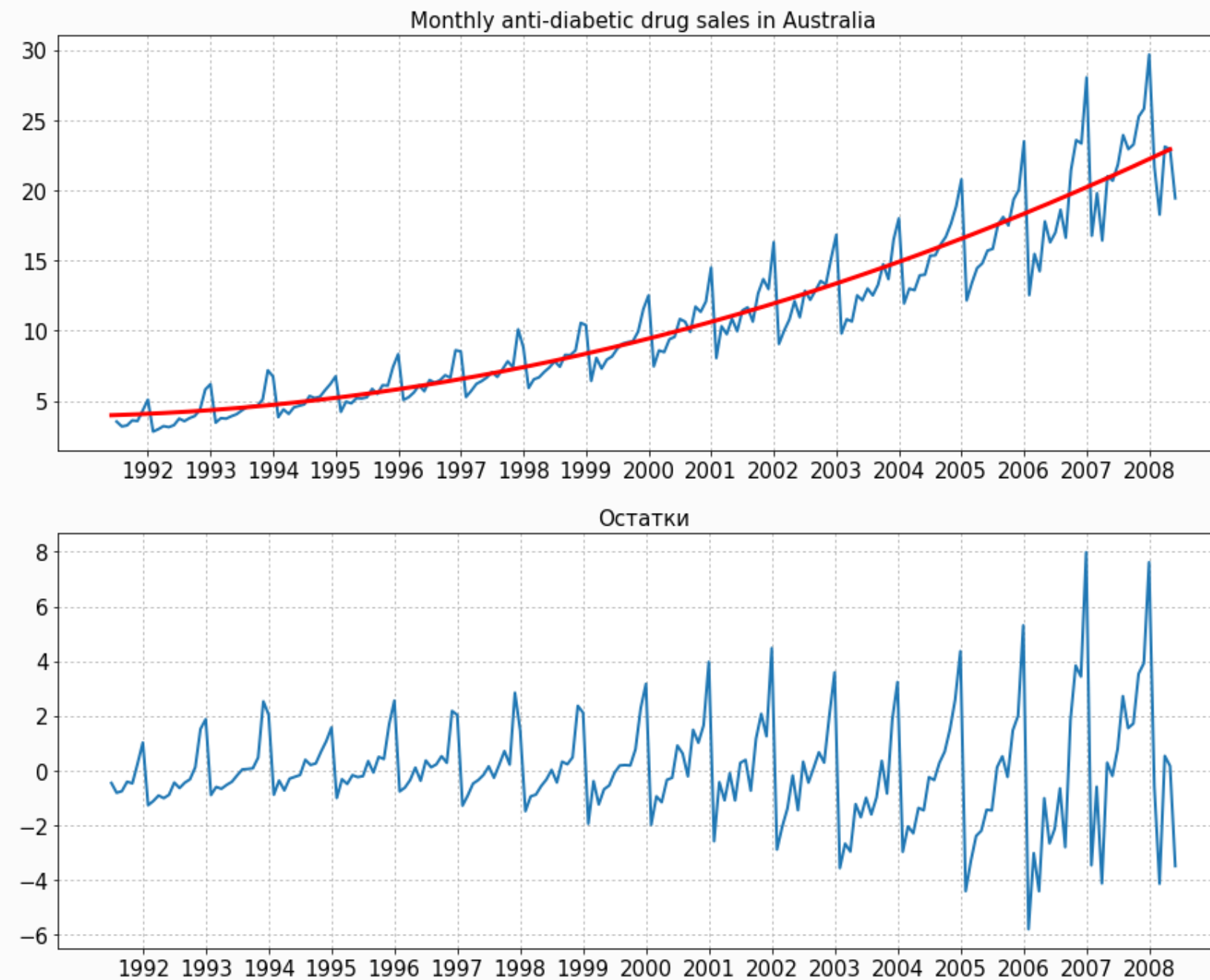
Цели темы

- ❶ Рассмотрите, как использовать классические модели регрессии для прогнозирования временных рядов
- ❷ Узнаете, как и зачем выделять фичи из временного ряда

Линейная регрессия



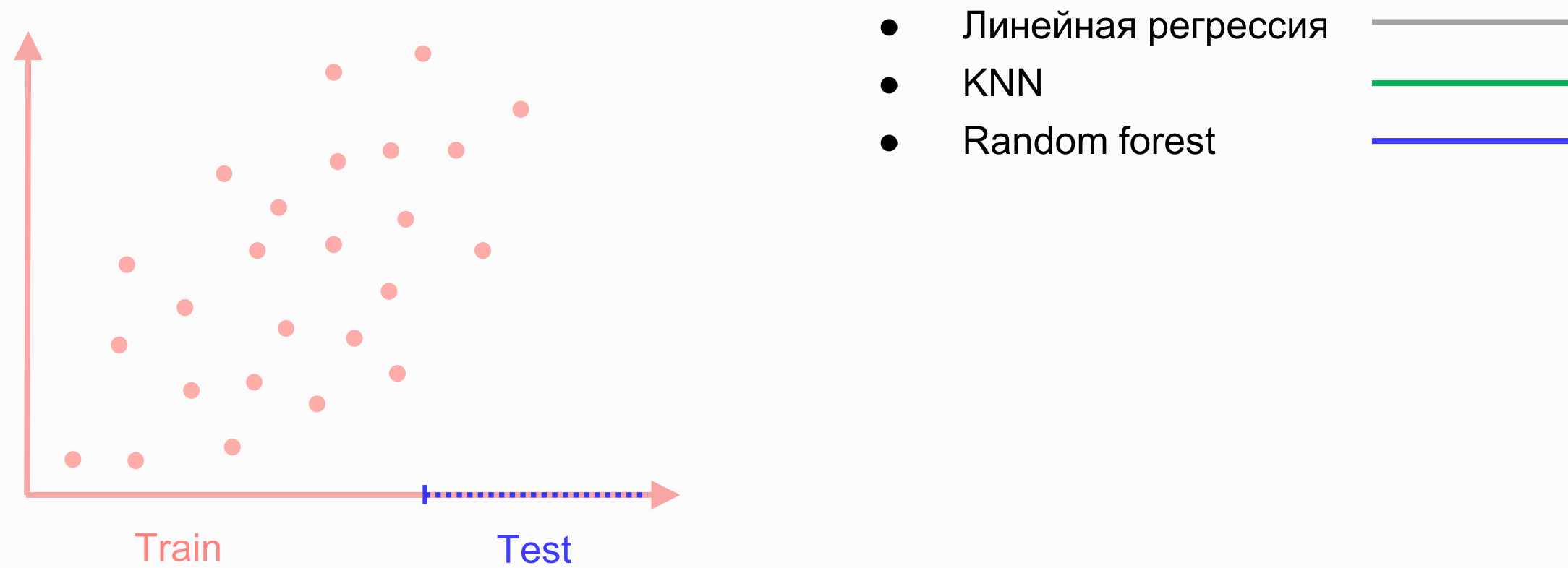
Добавление квадратичного признака



Задача с собеседования

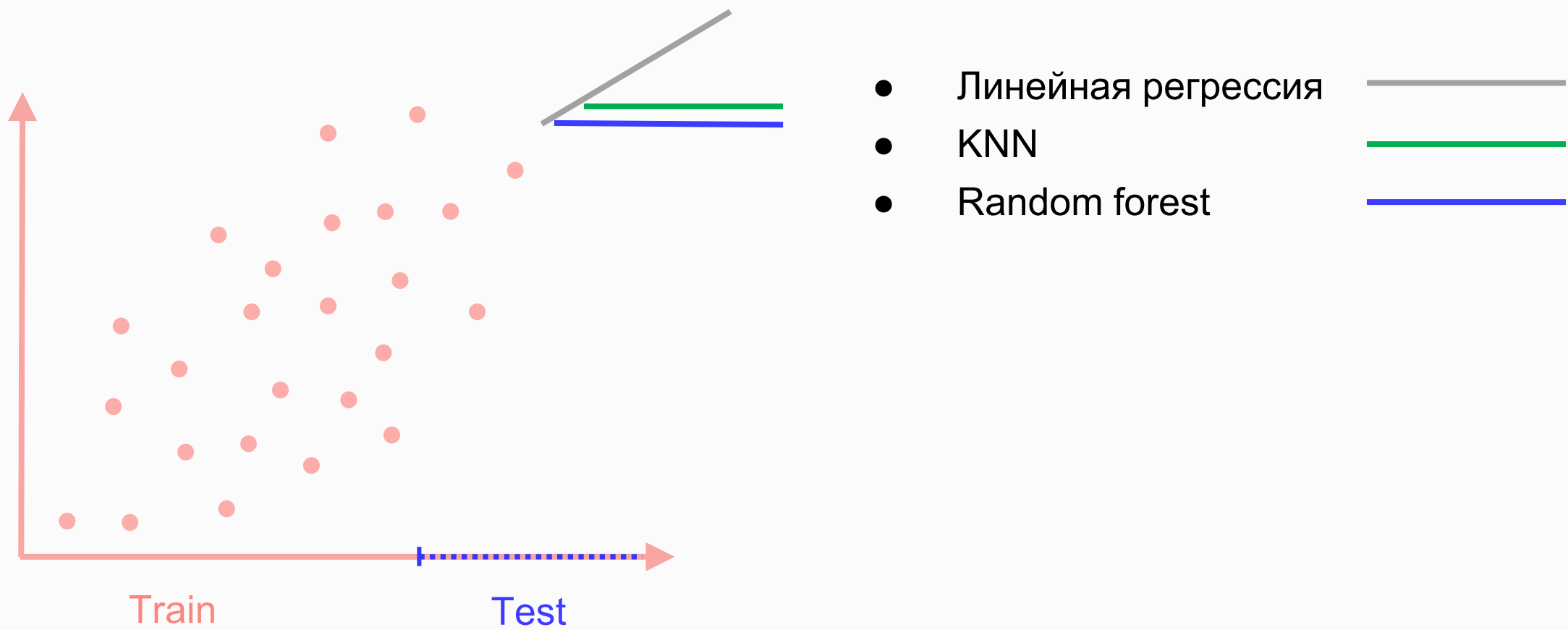
Есть три модели: линейной регрессии, KNN, random forest.

Как будут выглядеть предсказания на тесте?



Задача с собеседования: ответ

Высота линии в KNN-зависит от выбранного количества соседей.



Построение прогноза классическими моделями

Модель $y_t = f(y_1, \dots, y_{t-1})$,

где f — произвольная функция

Идея: построить функцию f некоторым ML-методом

ML-модели регрессии:

- линейная регрессия
- решающие деревья
- бустинги
- нейронные сети
 - свёрточные (CNN)
 - рекуррентные (RNN)

Построение прогноза классическими моделями

Возьмите предыдущие значения.

В день $60 * 24$ признака .

За полгода $60 * 24 * 180 \approx 260\,000$.

Вывод: для бустинга лучше не использовать значения временного ряда, а генерировать более разумные фичи.

Классические модели

Преимущества

- Удобство
- Много рядов — много моделей

Недостатки

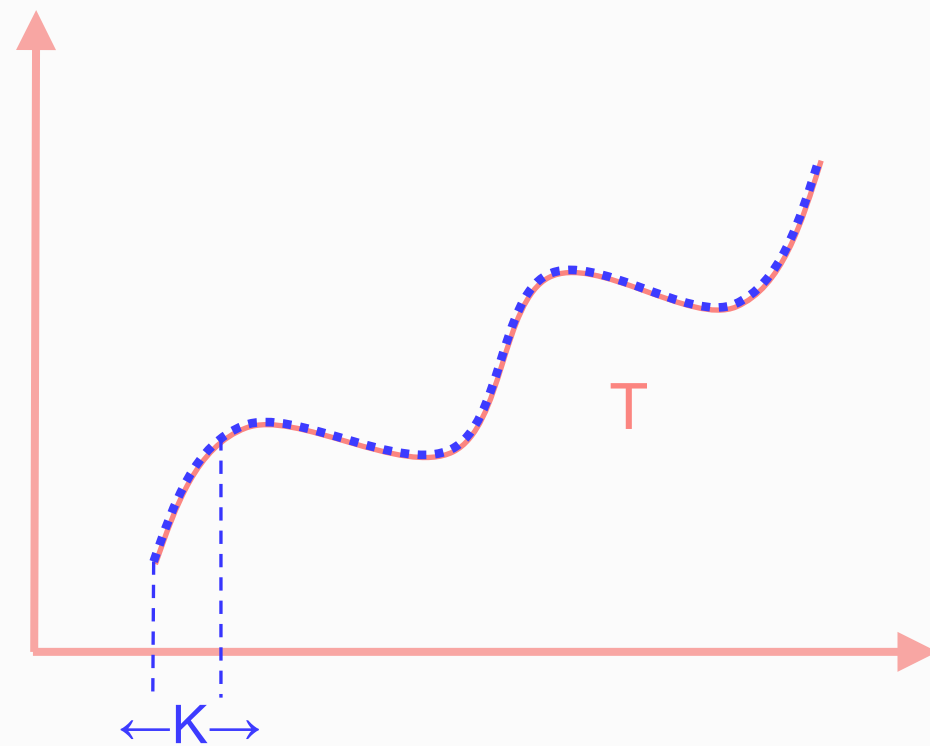
- Предсказательные интервалы не строятся напрямую
- Иногда работают хуже стандартных моделей
- Нужно обрабатывать признаки и генерировать фичи
- Понимание моделей может вызывать трудности

Примеры «рабочих» признаков для даты

- День недели
- Месяц
- Год
- Сезон
- Праздник
- Выходной
- Час

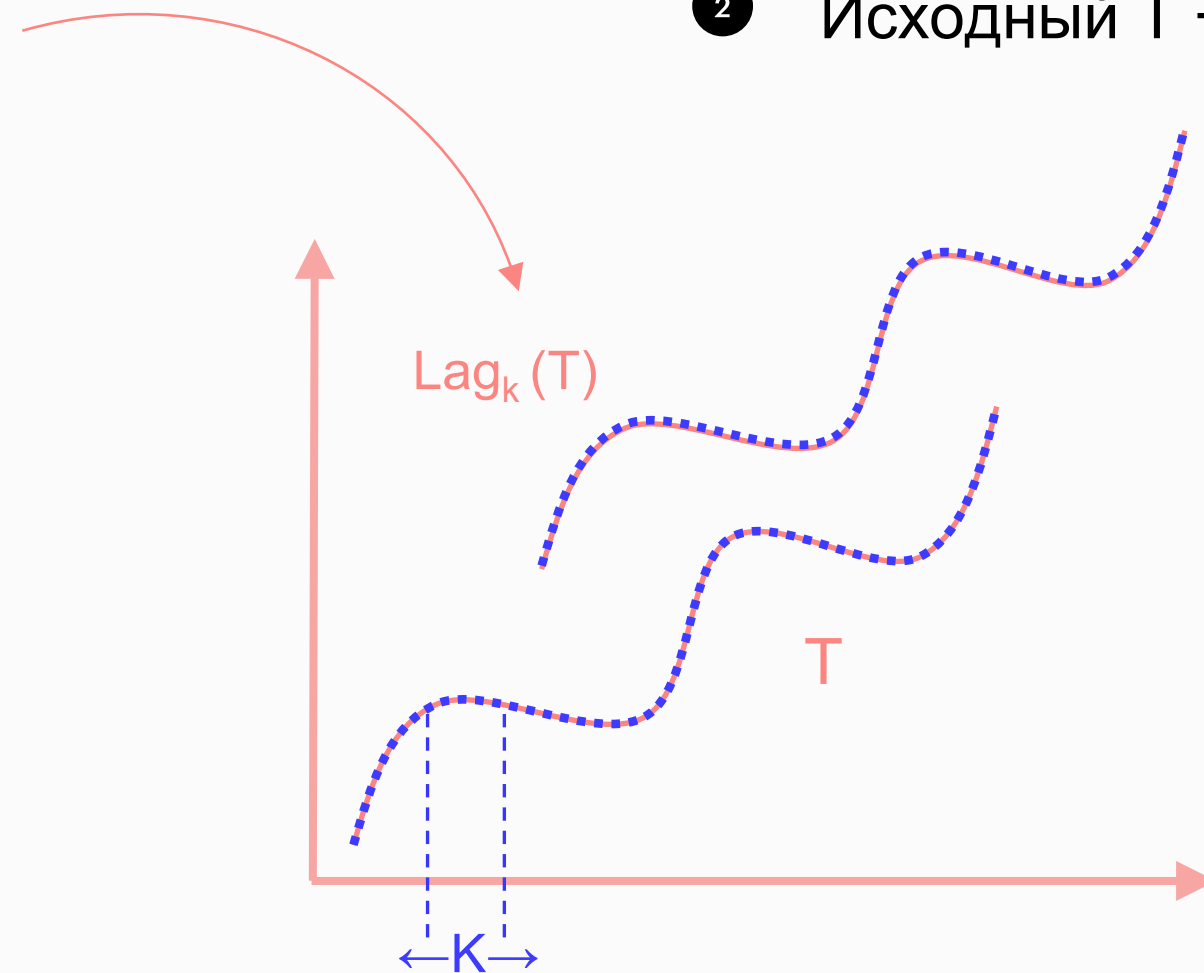
Предыдущее значение ряда (Лэг, Lag)

❶ Исходный ряд T



Алексей Подкидышев

❷ Исходный $T + \text{Lag}_k T$



Предыдущее значение ряда (Лэг, Lag)

Date	Value	Value _{t-1}	Value _{t-2}
1/1/2017	200	NA	NA
1/2/2017	220	200	NA
1/3/2017	215	220	200
1/4/2017	230	215	220
1/5/2017	235	230	215
1/6/2017	225	235	230
1/7/2017	220	225	235
1/8/2017	225	220	225
1/9/2017	240	225	220
1/10/2017	245	240	225

Главное о признаках

- Используйте только данные из прошлого
- Большое количество признаков может улучшить предсказание, но приведёт к вычислительным затратам
- Можно генерировать и другие признаки предметной области

Что делать, если после подсчёта статистик у вас появились None-ы?

Например, при подсчёте Lag N первые N значения ряда будут None.

Библиотека tsfresh

For extracting all features, we do:

1

```
from tsfresh import extract_features
extracted_features = extract_features(timeseries, column_id="id", column_sort="time")
```

You end up with the DataFrame *extracted_features* with more than 1200 different extracted features. We will now first, remove all NaN values (which were created by feature calculators that can not be used on the given data, e.g., because the statistics are too low), and then select only the relevant features:

2


```
from tsfresh import select_features
from tsfresh.utilities.dataframe_functions import impute

impute(extracted_features)
features_filtered = select_features(extracted_features, y)
```

Библиотека tsfresh

Библиотека ETNA

READMECode of conductLicense



Predict your time series the easiest way

pypi v2.7.1python 3.8 | 3.9 | 3.10Downloads 144k

coverage 10%tests passingdocs passinglicense Apache-2.0

channel telegramcontributors 31Stars 99

[Homepage](#) | [Documentation](#) | [Tutorials](#) | [Contribution Guide](#) | [Release Notes](#)

ETNA is an easy-to-use time series forecasting framework. It includes built in toolkits for time series preprocessing, feature generation, a variety of predictive models with unified interface - from classic machine learning to SOTA neural networks, models combination methods and smart backtesting. ETNA is designed to make working with time series simple, productive, and fun.

ETNA is the first python open source framework of [Tinkoff.ru](#) Artificial Intelligence Center. The library started as an internal product in our company - we use it in over 10+ projects now, so we often release updates. Contributions are welcome - check our [Contribution Guide](#).

Библиотека ETNA

GitHub, Inc / github.com

READMECode of conductLicense

Tutorials

We have also prepared a set of tutorials for an easy introduction:

Notebook	Interactive launch
Get started	launch binder
Backtest	launch binder
EDA	launch binder
Regressors and exogenous data	launch binder
Deep learning models	launch binder
Ensembles	launch binder
Outliers	launch binder
AutoML	launch binder
Clustering	launch binder
Feature selection	launch binder
Forecasting strategies	launch binder
Mechanics of forecasting	launch binder
Embedding models	launch binder
Custom model and transform	launch binder
Inference: using saved pipeline on a new data	launch binder
Hierarchical time series	launch binder
Forecast interpretation	launch binder
Classification	launch binder
Prediction intervals	launch binder
Working with misaligned data	launch binder

Выводы темы

- ✓ Разобрали задачу с собеседования об использовании регрессионных моделей для прогнозирования 2D-данных
- ✓ Изучили, как и зачем строить фичи при прогнозировании временного ряда регрессионными моделями
- ✓ Рассмотрели библиотеки для удобного построения фичей и прогнозирования рядов