СБЕР УНИВЕРСИТЕТ

Дипломный проект на тему:

«Прогноз расхода энергии по временным рядам»

Черепанова Ирина Николаевна

Актуальность темы

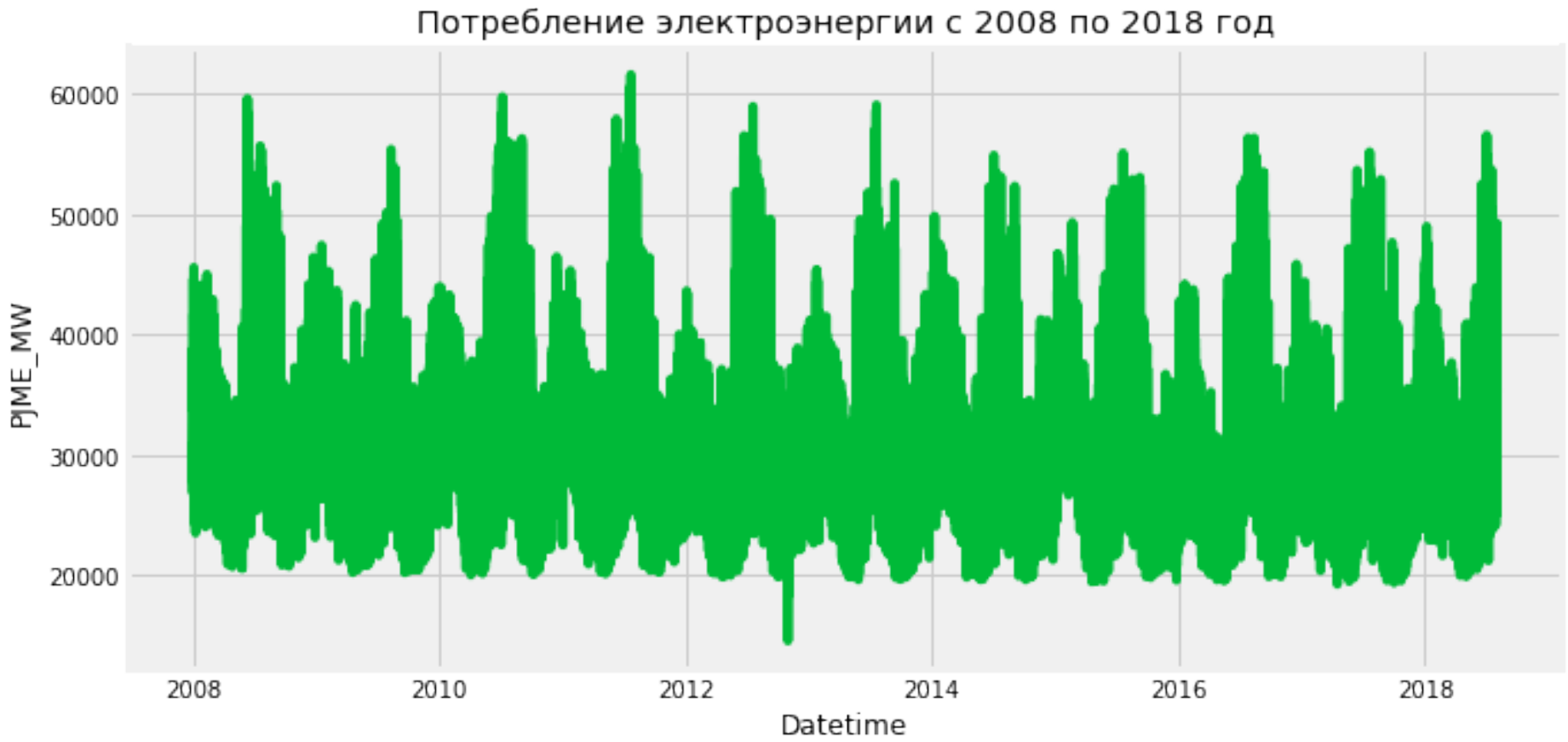
1. Прогнозирование имеет широкий спектр применений в различных областях.
2. Прогнозирование энергопотребления важно для планирования экономического развития городов и регионов.
3. При прогнозах важно использовать более точную модель.

Цель проекта:

1. Выявить в исследовании более точную модель для прогнозирования
2. Спрогнозировать потребление электроэнергии США с помощью выбранных моделей.

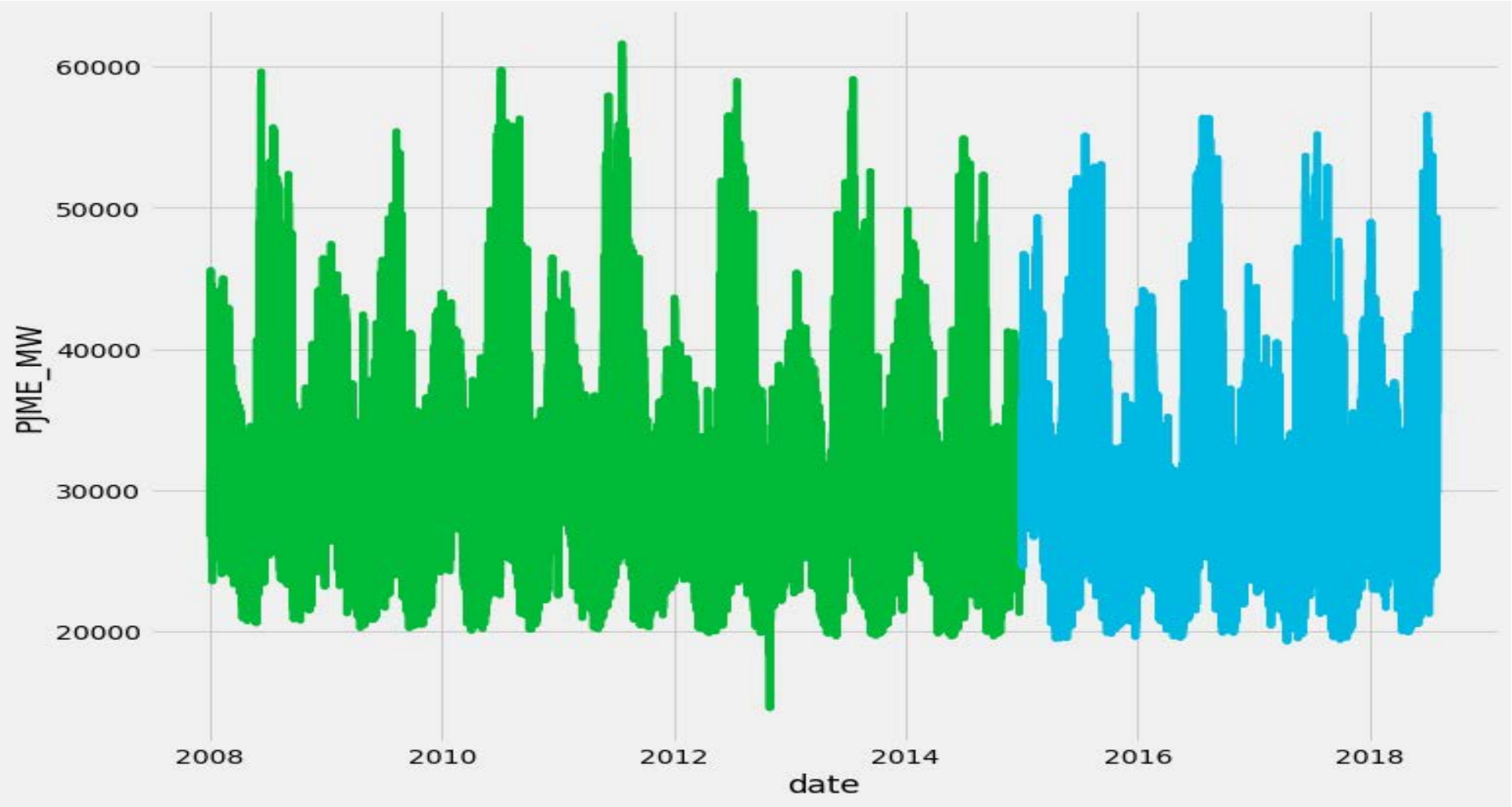
Черепанова Ирина Николаевна

Проведем анализ данных и построим временной ряд



Построим временной ряд по часам, дням недели, годам, неделям года

Построим временной ряд с train и test, чтобы визуализировать разделение.



Прогноз модели : с помощью Prophet

В основе этой методологии лежит процедура подгонки аддитивных регрессионных моделей (Generalized Additive Models, GAM) следующего вида:

$$y(t)=g(t)+s(t)+h(t)+\epsilon t, y(t)=g(t)+s(t)+h(t)+\epsilon t$$

где g(t) g(t) и s(t) s(t) — функции, аппроксимирующие тренд ряда и сезонные колебания (например, годовые, недельные и т.п.) соответственно, h(t)h(t) — функция, отражающая эффекты праздников и других влиятельных событий, а ϵtϵt — нормально распределенные случайные возмущения. Для аппроксимации перечисленных функций используются следующие методы:

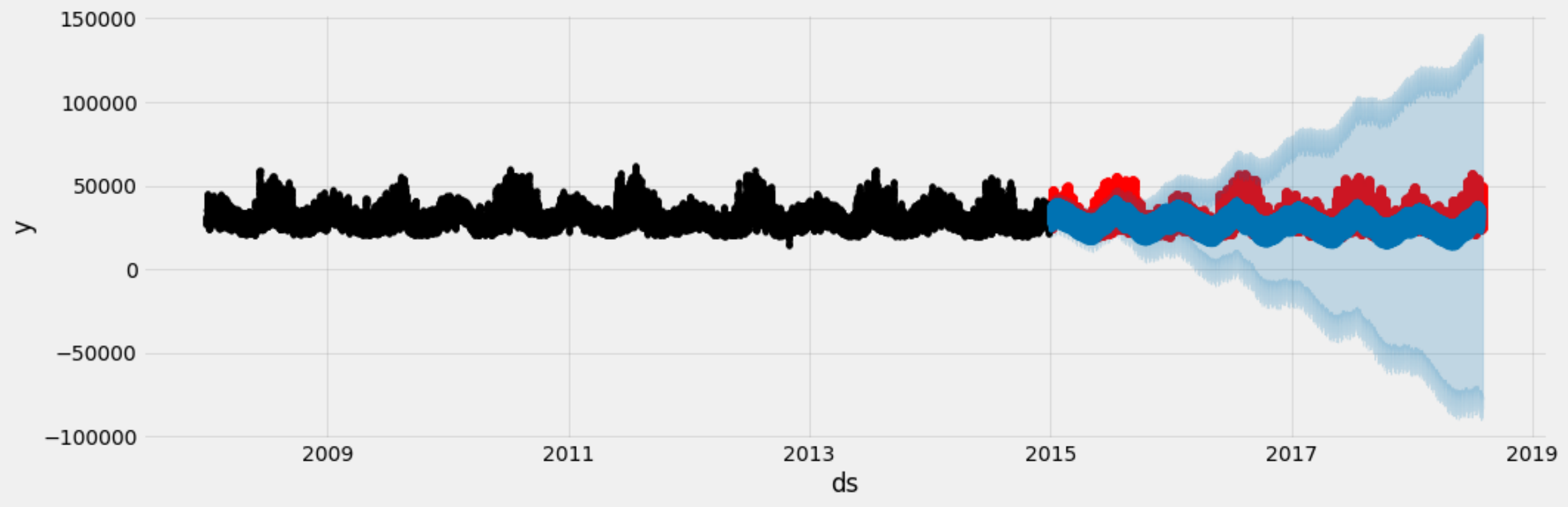
тренд: кусочная линейная регрессия или кусочная логистическая кривая роста;

годовая сезонность: частичные суммы ряда Фурье, число членов которого (порядок) определяет гладкость функции;

недельная сезонность: представлена в виде индикаторной переменной;

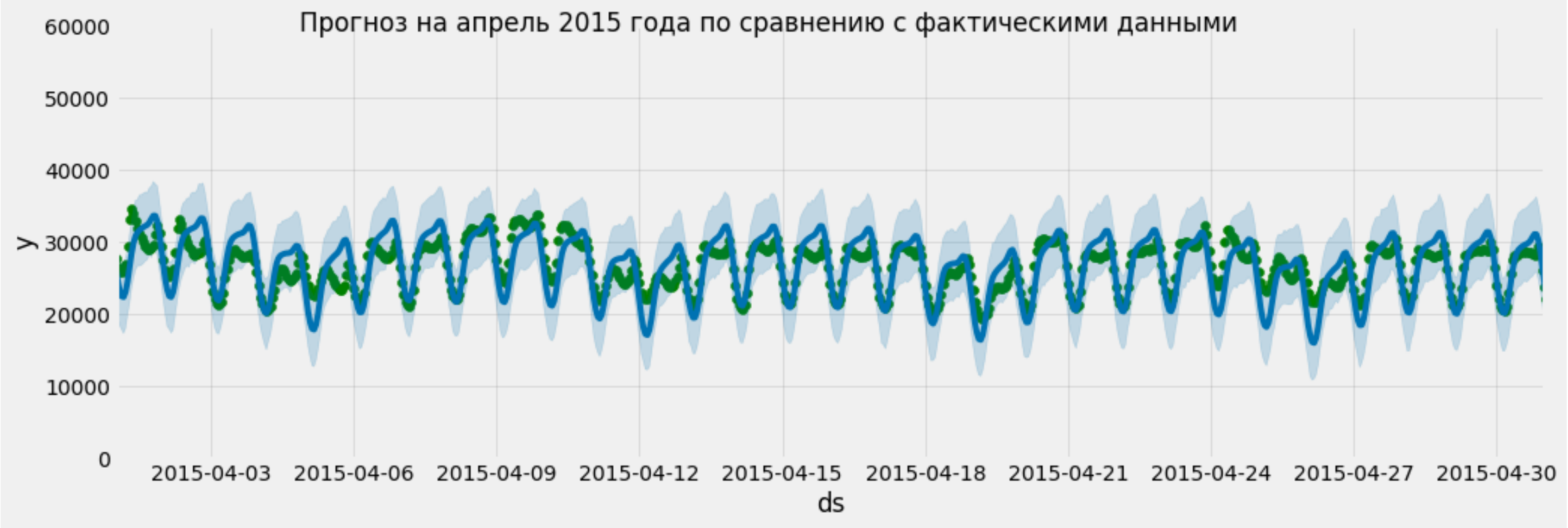
“праздники” (например, официальные праздничные и выходные дни — Новый год, Рождество и т.п., а также другие дни, во время которых свойства временного ряда могут существенно измениться — спортивные или культурные события, природные явления и т.п.): представлены в виде индикаторных переменных.
Оценивание параметров подгоняемой модели выполняется с использованием принципов байесовской статистики (либо методом нахождения апостериорного максимума (MAP), либо путем полного байесовского вывода). Для этого применяется платформа вероятностного программирования Stan.
Пакет prophet представляет собой ни что иное, как удобный интерфейс для работы с этой платформой из среды R

Сопоставим прогноз с фактическими данными



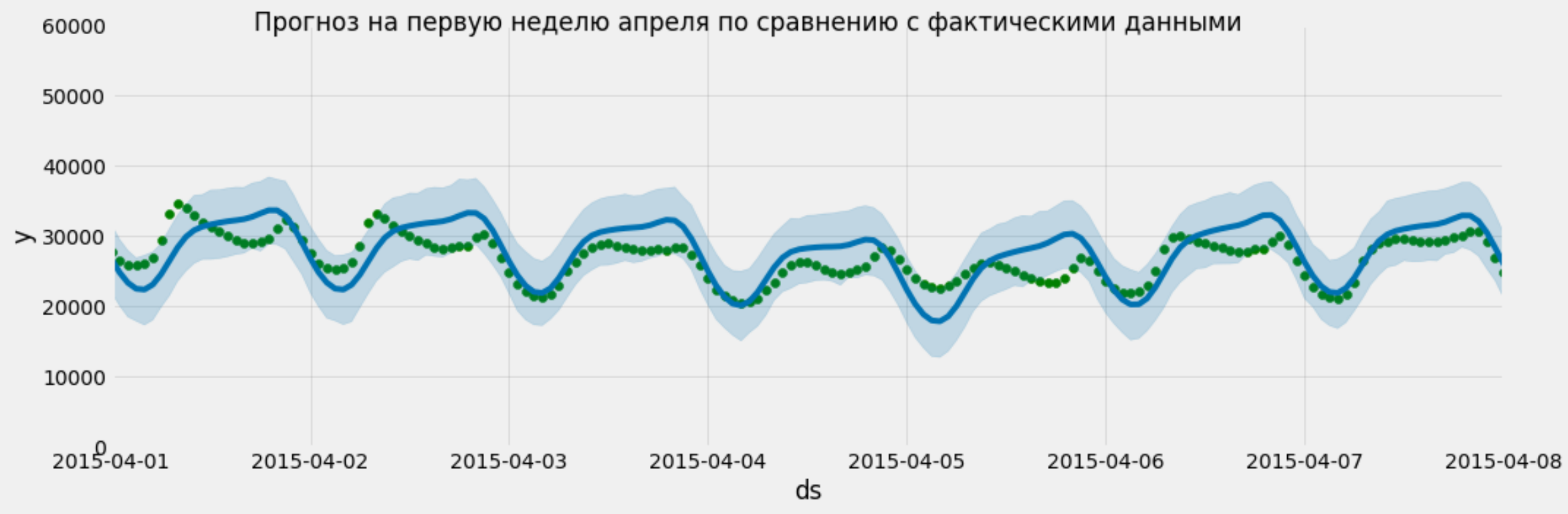
- Dark - это прошлые данные (черный цвет)
- Red - это фактические, актуальные
- Dark blue - это предсказание
- Light blue - интервал

Сравним на графике прогнозы на апрель месяц с реальными.



- Green - это фактические, актуальные
- Blue - это предсказание
- Light blue - интервал

Прогноз на первую неделю апреля по сравнению с фактическими данными



Показатели качества:

MSE на тестовой выборке: 48948325.21211523

MAE на тестовой выборке: 5464.6240741276315

MAPE на тестовой выборке : 16.96966125937518

Добавим праздники

Далее мы посмотрим, поможет ли добавление праздничных показателей точности модели

Метрики с учетом праздников

MSE на тестовой выборке: 49117710.88025762
MAE на тестовой выборке: 5477.6273848564915
MAPE на тестовой выборке : 17.009819878487757

Показатели качества без учета праздников

Показатели качества:
MSE на тестовой выборке: 48948325.21211523
MAE на тестовой выборке: 5464.6240741276315
MAPE на тестовой выборке : 16.96966125937518

Как видим, значения ошибок возросли, хоть и незначительно, в целом разница небольшая, но мы получили переобучение. Поэтому признак оказался неинформативным.

Черепанова Ирина Николаевна

Стоит учесть следующий факт периода: 29-30 октября ураган "Сэнди" вызвал сильные ветры и наводнения на значительной части восточной части Соединенных Штатов, в результате чего, по оценкам, 8 миллионов потребителей остались без электричества. Шторм, который обрушился на берег недалеко от Атлантик-Сити, штат Нью-Джерси, как ураган 1-й категории, в конечном итоге оставил без электричества множество домов и предприятий в Нью-Джерси (2,7 миллиона), Нью-Йорке (2,2 миллиона), Пенсильвании (1,2 миллиона), Коннектикуте (620 000), Массачусетсе (400 000), Мэриленде (290 000), Западной Вирджинии (268 000), Огайо (250 000) и Нью-Гемпшире (210 000). Также сообщалось о перебоях в подаче электроэнергии в ряде других штатов, включая Вирджинию, Мэн, Род-Айленд, Вермонт и округ Колумбия. Поэтому, необходимо это учесть в нашем исследовании и стоит почистить данные от неверных показателей.

Показатели качества с учетом чистых данных

Показатели качества:
MSE на тестовой выборке: 48839822.7037
MAE на тестовой выборке: 5457.1816
MAPE на тестовой выборке : 16.9490
Вывод:
После очистки данных мы видим что показатели улучшились, хоть и не намного. Можно сказать что после очистки мы видим небольшое улучшение показателей по сравнению с первой моделью. Дополнительная очистка данных наряду с праздниками показывает немного лучшие результаты, что и показали данные метрики.

Прогноз модели : с помощью SARIMAX

ARIMA и SARIMAX

ARIMA (англ. autoregressive integrated moving average)- интегрированная модель авторегрессии — скользящего среднего — модель и методология анализа временных рядов.

Авторегрессионная модель (AR, autoregressive model) — модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда.

Модель скозящего среднего (MA, moving average model) - модель, в которой моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, т.е. разностей между прошлыми фактическими и теоретическими уровнями.

ARIMA(p, d, q) = AR(p) + MA(q) + I(d) , где I(k) - интегрируемый ряд порядка k

SARIMAX - модель временных рядов, построенная на основе расширенной (eXtended) модели ARIMA с добавлением сезонности (Seasonal).

Учитывая данные временного ряда X_t, где t — целочисленный индекс, а X_t — действительные числа, $\mathrm{ARMA}(p', q)$ модель предоставлена

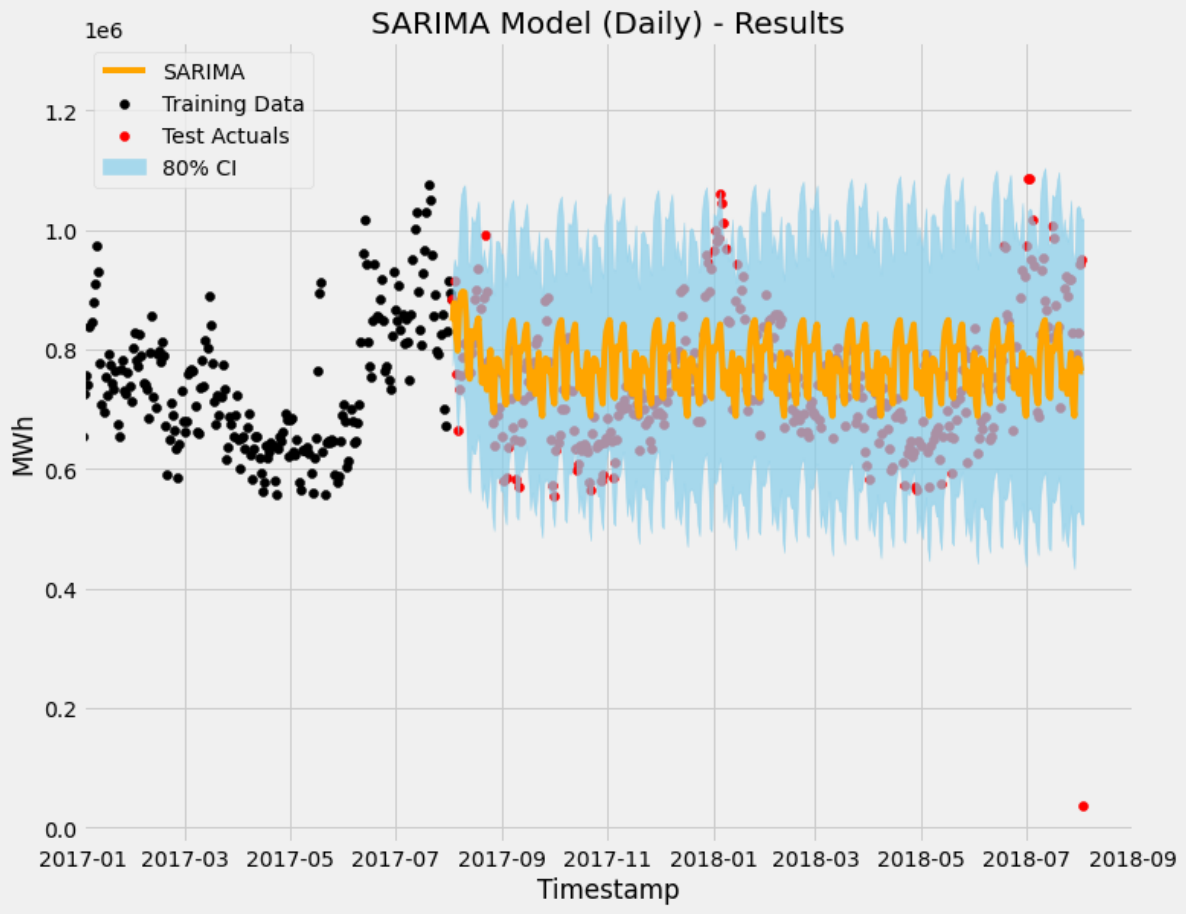
$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

или эквивалентно

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

Прогноз модели : с помощью SARIMAX

Несезонные модели ARIMA обычно обозначаются ARIMA(p , d , q), где параметры p , d и q являются неотрицательными целыми числами, p — порядок (количество временных задержек) авторегрессионной модели , d — степень разность (количество раз, когда из данных вычитались прошлые значения), а q — порядок модели скользящего среднего . Сезонные модели ARIMA обычно обозначаются ARIMA(p , d , q)(P , D , Q)$_m$, где m относится к количеству периодов в каждом сезоне, а прописные буквы P , D , Q относятся к авторегрессионному, разностному и скользящему среднему для сезонной части модели ARIMA. [8] [2]



Метрики качества

Среднеквадратичная ошибка RMSE: 23648465438.8249

Средняя абсолютная ошибка MAE:

131987.15927191788796335459

Средняя абсолютная ошибка в процентах MAPE: 24.94

Прогноз модели : с помощью LASSO (L1) Regression

представляет собой метод регрессионного анализа , который выполняет как выбор переменных , так и регуляризацию , чтобы повысить точность прогнозирования и интерпретируемость результирующей статистической модели ..

Лассо изначально был разработан для моделей линейной регрессии. Этот простой случай позволяет многое узнать об оценщике. К ним относятся его связь с гребневой регрессией и выбором наилучшего подмножества , а также связь между оценками коэффициентов лассо и так называемой мягкой пороговой обработкой. Это также показывает, что (как и при стандартной линейной регрессии) оценки коэффициентов не обязательно должны быть уникальными, если ковариаты коллинеарны .

Хотя изначально регуляризация лассо была определена для линейной регрессии, она легко распространяется на другие статистические модели, включая обобщенные линейные модели , обобщенные оценочные уравнения , модели пропорциональных рисков и М-оценки

Регрессия ЛАССО включает регуляризацию и выбор признаков в свой алгоритм. Регуляризация - это метод, используемый в регрессионных алгоритмах, чтобы избежать переобучения.

Прогноз модели : с помощью LASSO (L1) Regression

1. LASSO (L1) Regression

Среднеквадратичная ошибка RMSE: 92173.65044531054

Средняя абсолютная ошибка MAE: 62908.35600470739

Средняя абсолютная ошибка в процентах MAPE: 14.50

Чем меньше ошибка, тем лучше / точнее модель, поэтому в случае оценки по MAPE лучше показала модель LASSO (L1) Regression.

Черепанова Ирина Николаевна

15

Оценка моделей с использованием метрик среднеквадратической ошибки (RMSE), средней абсолютной ошибки (MAE), , средняя абсолютная ошибка в процентах (MAPE)

1. МОДЕЛЬ Prophet

Среднеквадратичная ошибка RMSE: 48839822.7037

Средняя абсолютная ошибка MAE: 5457.1816

Средняя абсолютная ошибка в процентах MAPE: 16.9490

1. МОДЕЛЬ SARIMAX

Среднеквадратичная ошибка RMSE: 23648465438.8249

Средняя абсолютная ошибка MAE: 131987.15927191788796335459

Средняя абсолютная ошибка в процентах MAPE: 24.94

1. LASSO (L1) Regression

Среднеквадратичная ошибка RMSE: 92173.65044531054

Средняя абсолютная ошибка MAE: 62908.35600470739

Средняя абсолютная ошибка в процентах MAPE: 14.50

Выводы

Анализируя показатели оценки качества, стоит сказать следующее:
Чем меньше ошибка, тем лучше / точнее модель, поэтому в случае оценки по MAPE лучше показала модель LASSO (L1) Regression. Тем не менее, Prophet занимает хорошие позиции по метрикам MAE и RMSE
Мы предсказали потребление энергии, исследовали методы анализа временных рядов, выявили лучшие модели. Цель проекта можно считать выполненной.

Список использованных источников

1. https://habr.com/ru/company/vk/blog/513842/
2. https://en.wikipedia.org/wiki/Lasso_(statistics)
3. https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html
4. https://habr.com/ru/company/ods/blog/323730/