Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра Системного Программирования

Гарифуллин Шамиль Раифович

Прогнозирование потребления электроэнергии с помощью методов машинного обучения

Бакалаврская работа

Допущена к защите.
Зав. кафедрой:
д.ф.-м.н., профессор Терехов А. Н.

Научный руководитель:
д.ф.-м.н., профессор Терехов А. Н.

Рецензент:
д.ф.-м.н., доцент Графеева Н. Г.

Санкт-Петербург
2015

SAINT PETERSBURG STATE UNIVERSITY
Mathematics & Mechanics Faculty

Software Engineering Chair

Garifullin Shamil

Prediction of energy consumption via Machine Learning methods

Bachelor's Thesis

Admitted for defence.
Head of the chair:
professor Andrey Terekhov

Scientific supervisor:
professor Andrey Terekhov

Reviewer:
associate professor Natalia Grafeeva

Saint Petersburg
2015

Оглавление

Введение

Методы машинного обучения имеют обширную область практических применений в автоматике, управлении, экономике, социологии, медицине, геологии, астрономии, ядерной физике и т.д. Машинное обучение включает в себя задачи различного спектра: от построения графиков до оценки качества данных. Основной проблемой области является использование без достаточного понимания методов и техник машинного обучения, данная работа была проведена с целью показать возможности и особенности методов области.

Эффективное энергопроизводство и энергопотребление является одной из важнейших экологических проблем современности. В особенности в ситуации, когда основным источником производства является газовая энергия, используемая в ТЭЦ.

В данной работе рассмотрены способы прогнозирования потребления электроэнергии промышленными предприятиями города Санкт-Петербург. Одним из требований эффективного прогнозирования являлось точное предсказания потребления электроэнергии на каждый час последующих суток для оптимального распределения ресурсов этих предприятий.

Также целью работы являлось изучение возможности предсказания такого рода при помощи данных о погоде на то же время, что и замеры потребления электроэнергии, в виду сильной зависимости этих данных.

Задачу потребления электроэнергии можно решать двумя способами, для этого её можно сформулировать следующим образом:

- Задача регрессии — одна из наиболее распространенных задач в анализе данных и создании предсказаний. Для решения этой задачи требуется создание функции регрессора, которая присваивает

каждому набору входных атрибутов вещественное значение. Регрессия входных значений производится после прохождения этапа ‘обучения’, в процессе которого на вход обучающего алгоритма подаются входные данные с уже приписанными им значениями результатов.

- Задача прогнозирования временных рядов — одна из наиболее распространенных задач прогнозирования в прикладной индустрии и актуарной математике. Для решения этой задачи используют методы выделения закономерностей в данных, таких как: сезонность, тренд и помехи. Также используются различного рода моделирования физических, стохастических и прочих процессов.

Для обеих задач (в особенности задачи регрессии) нужно правильно обработать данные и зафиксировать методы оценивания полученных результатов. В качестве методов оценки были применены как различного рода метрики, так и графические подходы для визуализации отклонений от истинных значений. В данной работе рассмотрены различные подходы к обработке данных.

1. Постановка задачи

В рамках данной работы ставились следующие задачи:

- Показать возможность прогнозирования почасового потребления электроэнергии с использованием данных о погоде для дальнейшего анализа другими участниками работы
- Найти и исправить несоответствия в данных
- Заполнить пробелы в данных о погоде
- Выделить новые параметры, влияющие на результат
- Ввести критерии оценки прогнозирования

2. Описание особенностей данных

Данные, на основе которых строится модель, имеют в разные форматы и получены из различных источников.

Ниже представляется описание данных.

2.1. Данные об энергопотреблении

Формат данных имеет вид **.xml** [23] документа, который получен с помощью запроса к **Oracle DBMS** [15]. Данные аггрегированы по часам. Измерения проведены за каждый час 2009 и 2010 года.

Ниже приведен пример входных данных об электричестве.

Listing 1: Потребление электроэнергии

```
<?xml version="1.0" ?>
<!DOCTYPE main [
  <!ELEMENT main (DATA_RECORD*)>
  <!ELEMENT DATA_RECORD (VALUE?,DATE?,HOUR?)+>
  <!ELEMENT VALUE (#PCDATA)>
  <!ELEMENT DATE (#PCDATA)>
  <!ELEMENT HOUR (#PCDATA)>
]>
<main>
  <DATA_RECORD>
    <VALUE>243887</VALUE>
    <DATE>01.01.2009 0:00:00.000000</DATE>
    <HOUR>1</HOUR>
  </DATA_RECORD>
  <DATA_RECORD>
    <VALUE>243068</VALUE>
    <DATE>01.01.2009 0:00:00.000000</DATE>
    <HOUR>2</HOUR>
  </DATA_RECORD>
</main>
```

Как видно это xml файл с фиксированной схемой данных, что является распространённой практикой в мире СУБД.

2.2. Данные о погоде

Данные о погоде аггрегированы из нескольких источников и включают в себя такие переменные как:

- Дата измерения
- Час измерения
- Температура
- Влажность
- Точка росы
- Атмосферное давление
- Направление ветра
- Скорость ветра
- Процент облачности
- Высота облаков
- Видимость
- Погодные аномалии такие, как снег, туман, дымка, дождь и т.д.

Основным отличием от данных об энергопотреблении являются замеры, делавшиеся каждые три часа (в отличие от каждого часа).

Формат данных имеет вид **.xslx** [24] документа заполненного от руки с двенадцатью листами, то есть разделением на каждый месяц. Измерения также за 2009 и 2010 года.

2.3. Несоответсвия в данных и пробелы в данных о погоде

Так как погодные данные заполнены от руки, в них могут закрасться опечатки и сложно машинноинтерпретируемые условности, такие как семантически сходные понятия и определения, которые являются близкими, но не тождественными. Также, стоит отметить, что из-за конфиденциальности предприятий данные о погоде не могут быть получены из открытых источников, поэтому проведен дополнительный анализ данных.

Так как наиболее известные и изученные алгоритмы из области машинного обучения нуждаются в соответсвии прогнозируемых данных данным на основе которых происходит прогнозирование, также необходимо провести интерполяцию для возможности применения известных алгоритмов регрессии.

2.4. Новые параметры

Как правило в задачах анализа данных доступные переменные могут не содержать всей информации, которая нужна для успешного прогнозирования результатов. Но сами данные могут содержать скрытые переменные, при выявлении которых результаты анализа могут радикально улучшиться [6][13]. В данной работе примером такой переменной может послужить месяц произведенных замеров.

Также не стоит забывать о переменных которые можно добавить извне [12]. Конкретно в этой работе можно выделить такие переменные как Праздничные дни, HS index и Humidex Index.

3. Обзор аналогичных задач и способов их решения

3.1. На основе данных о домах

Одной из аналогичных работ является задача предсказания потребления электроэнергии на основе данных о 1500 домах Гонг-Конга [9].

В работе были собраны данные обо всех энергопотребляющих приборах домов и также сделаны пометки о возрасте и прибыли людей, проживающих в этих домах. Измерения делались ежедневно, проведены как зимой, так и летом.

После обработки проведен анализ данных и построены три модели прогнозирования энергопотребления:

- **Линейная регрессия** [26]
- **Нейронные сети** [20]
- **Деревья решений** [16]

Как видно из таблицы ниже лучший результат дают **Нейронные сети** и **Деревья решений** с отклонением от истинного значения в 5-6%.

	Результаты
Линейная регрессия	7.6-8.2%
Нейронные сети	5.5-6.7 %
Деревья решений	4.8-5.6%

3.2. Только на основе данных об энергопотреблении

В другой работе использованы иные методы анализа данных и прогнозирования результатов, в первую очередь из-за того, что данные в этой задаче имеют другой вид [22]. А именно: ежесекундный отчет о потреблении электроэнергии из каждого узла двух кластеров компьютеров, состоящих из 56 и 79 узлов. Что результировало в большом объеме данных (порядка 50 Гб).

Для обработки такого количества информации были использованы **Hadoop** [1] и **Pig** [2], развернутые на кластере из 55 узлов. Hadoop был выбран для удобного аггрегирования данных по часам с использованием технологии **Mapreduce** [7]. А Pig, являясь надстройкой над Hadoop для выполнения запросов на **SQL**-подобном [19] языке, использовался для формирования Mapreduce запросов.

Данные были аггрегированы представленным ниже способом:

```
x = LOAD '$in' USING PigStorage(',')
y = FILTER x BY watts>=0;
z = FOREACH y GENERATE
    node, watts;
g = GROUP z BY (node_ID);
s = FOREACH g GENERATE flatten(group),
    COUNT(z), MIN(z.watts),
    MAX(z.watts), AVG(z.watts),
STORE s INTO '$out' USING PigStorage(',');
```

Затем применен стандартный метод анализа временных рядов **ARIMA** [3], который выделяет шум, общую тенденцию временного ряда и периодичность в отдельные компоненты, для построения упрощенной модели временного ряда.

В этой работе стоит отметить малый набор факторов, использованный в анализе. Однако получены качественные результаты. Можно предположить, что, в первую очередь, это обусловлено высокой частотностью сбора показаний. В итоге получено отклонение от истинного значения в 8-9%.

3.3. На основе данных о домах и погоде

Ещё одной работой, которую стоит рассмотреть — является задача предсказания потребления электроэнергии на основе данных о погоде, жителях и домах района города Палермо [21].

Были собраны данные об энергопотреблении домов и данные о погоде в том же регионе. Измерения проводились каждый час.

После обработки данных в работе также посчитаны новые переменные на основе старых:

- **Индекс Humidex** [10] — наиболее популярная мера дискомфорта погодных условий.
- **Индекс HS** [21] — Индекс Humidex с учетом вероятности включения термостата в жилище, в работе моделировался вероятностно[1].

Индекс Humidex может быть вычислен по следующей формуле:

$$H = T + 5/9 \cdot (e - 10).$$

Где e — давление жидкости в воздухе, которое может быть оценена по следующей формуле:

$$e = 6.112 \times 10^{[(7.5 \cdot T)/(237.7+T)]} \cdot RH/100.$$

Где RH — влажность воздуха.

[1]Вероятность включенного в квартире термостата в момент времени t моделировалась сигмоидальной функцией $P_t = (1 + e^{-0.210317 \cdot (H_t - 41)})^{-1}$.

Моделью прогнозирования служила **Рекуррентная Нейронная Сеть** [11] — глубокая Нейросеть, обладающая свойством памяти и часто использующаяся в анализе временных явлений.

В результате получено отклонение от истинного значения в 4%.

4. Обработка данных

4.1. Приведение к стандартному виду

Сперва данные приводились к единому формату — таковым выбран формат **.csv** [5], поэтому была разобрана xlsx книга о погодных данных и соединена в один файл csv.

Так как информация о погоде заполнена от руки, в ней могли потенциально обнаружиться несоответствия. С помощью среды анализа данных **Rstudio** [8] и языка **R** [18] найдены отсутствующие измерения и вручную внесены исправления. Некоторые значения погодных аномалий были заполнены без использования определенного формата (например "Дождь" - "Дождливая погода") — соответсвенно внесены поправки в такие данные, путем выбора единого названия для семантически сходных понятий. Также построены гистограммы значений с целью выявить аномалии. Аномалий не обнаружено.

Данные об электричестве предоставлены в формате xml — считаны и записаны в формат csv. Крупных отклонений не обнаружено, изменения не вносились.

Наиболее крупным несоответствием данных формату являлся переход на летнее время в погодных данных, что устранено приведением времени к единому временному формату без переходов на летнее время.

4.2. Новые переменные

Основным источником новых переменных в данной работе является отметка о времени. Она разделена на несколько независимых переменных, таких как:

- День недели

- Месяц
- Номер недели в году

Из работы [21] взята переменная Индекс Humidex.

Также выделен новый параметр ‘Праздники’, заполненный с помощью каллендарных праздников РФ.

4.3. Интерполяция данных

В следствие того, что информация о погоде дана на каждые три часа, а прогнозирование делается на каждый час последующего дня, принято решение проинтерполировать данные.

Для категориальных переменных, таких как направление ветра и атмосферные аномалии интерполяция проведена следующим образом: для данного, предыдущего и последующего часов принимаются одинаковые значения.

Для численных переменных рассмотрены различные методы интерполяции [25], так как результаты не дали разницы в прогнозировании выбрана линейная интерполяция, как способ интерполированя численных данных.

5. Методы оценки

5.1. Метрика R^2

Метрика R^2 [17], также известная как Коэффициент Детерминации. Вычисляется по следующей формуле:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Метрика выбрана в первую очередь из-за того, что является нормированным показателем качества прогнозирования модели и позволяет сравнивать как модели между собой, так и непосредственно делать предположения о качестве результатов. Метрика оценивает количество объясненной вариации модели.

5.2. Графики отклонений

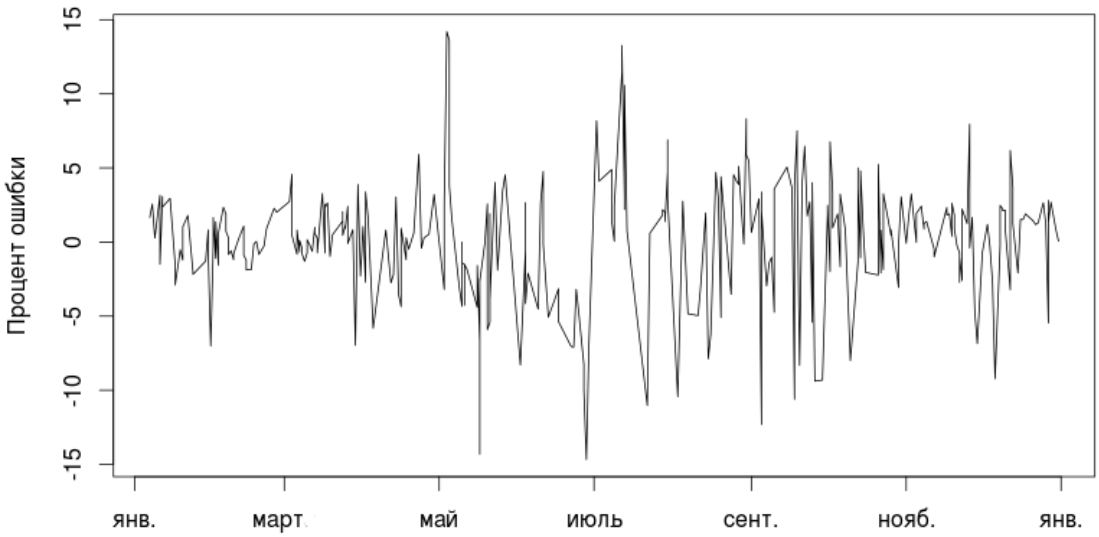

Рис. 1: Результаты за 2010 год

Также, чтобы не полагаться лишь на один одномерный показатель,

принято решение строить графики отклонений для каждого выбранного способа, как, например, видно на графике для деревьев решений (график 5.2).

6. Методы прогнозирования

6.1. Наивный метод

Изначально, чтобы получить базовый результат, который можно затем улучшать, испробован наивный способ прогнозирования.

Он заключается во взвешеном прогнозировании результатов на конкретный час с использованием данных за предыдущие дни с различного рода весами.

После программатического перебора нескольких вариантов лучший результат показало экспоненциальное взвешивание потребления за предыдущие три дня, а именно — тот же час предыдущего дня с весом 0.7, за день до этого с весом 0.25 и еще один день с весом 0.05.

Получен результат с метрикой $R^2 = 0.912$

6.2. Деревья решений

Деревья решений — жадный алгоритм построения цепочки if-then-else (график 6.2) для построения моделей прогнозирования данных. Позволяет получать стабильные решения, сравнимые с SVM и Нейросетями, при этом не используя больших вычислительных мощностей, требуемых предыдущими.

Оценка результатов проведена методом **Bootstrapping**'а [4]. А именно, берется множество выборок (в данном случае 100) на них каждый раз заново обучается дерево решений и прогнозируется следующий день по каждому часу. Сами множества выборок меняются — это отрезки времени длиной от трёх до двенадцати месяцов, с началом в случайный момент времени.

В итоге получены достаточно стабильные результаты, даже при варьировании длины промежутка обучения.Результат с метрикой $R^2 = 0.974 \pm 0.009$

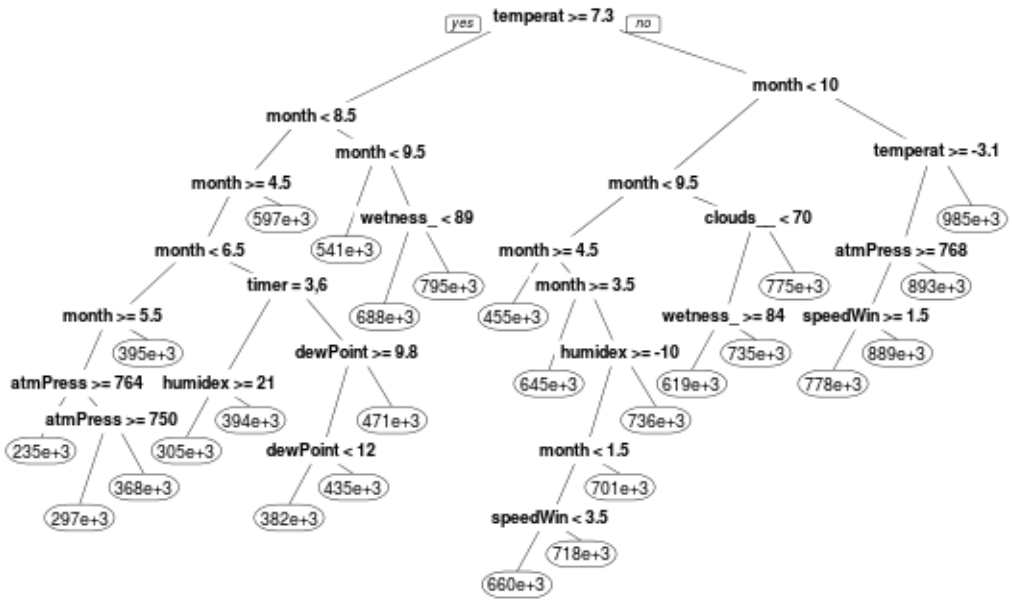

Рис. 2: Одно из построенных деревьев решений (обрезано для читаемости)

6.3. Лес решений

Лес решений — это модифицированный алгоритм деревьев решений, который строит множество деревьев на различных подмножествах обучающей выборки и, в следствие **Закона Больших чисел** [14], получает лучшие результаты путем выбора среднего из всех прогнозов деревьев.

Обучение и апробация проведено такими же методами, как и для деревьев решений.

В итоге получены стабильные результаты, даже при варьировании длины промежутка обучения от трёх до двенадцати месяцев. Получен результат с метрикой $R^2 = 0.981 \pm 0.004$

7. Влияние новых переменных на результаты

Алгоритм Деревьев Решений позволяет выделить параметры, оказывающие наибольшее влияние на силу алгоритма. Это делается путем удаления переменных, для которых производится оценка важности и рассмотрения разницы полученного отлонения от изначального. Ниже представлена эта самая разница в процентном соотношении.

	t, C	Т. Росы	Месяц	Облач.-ть	Д. Нед.	Празд.[2]
Сниж.-е точности	23%	18.5%	16.7%	6.5%	5.2%	0.4% (2.2%)

Из полученных результатов можно сделать следующие выводы:

- Новые параметры, полученные из данных, хоть и не являются самыми важными, однако вносят значительный вклад в силу модели
- Параметр же ‘Праздники’, полученный извне, не является важным для модели в целом, но имеет некоторое влияние на дни, объявленне праздниками РФ.

[2]В скобках указан результат только для праздников.

Заключение

В ходе работы были изучены и апробированы различные подходы к анализу и обработке данных. Показаны возможные способы их применения.

Также изучены и использованы метрики и визуализации для возможности ведения правильного анализа данных и оценки результатов.

Проведена стандартизация и преобразование данных к единому виду. Выделены новые переменные из уже существующих и добавлены новые внешние переменные. Перечислены возможные способы последующего анализа и улучшения результатов данной работы.

Показана возможность эффективного прогнозирования почасового потребления электроэнергии, как на основе предыдущих измерений, так и с использованием погодных данных. Что позволит в дальнейшем с уверенностью решать как похожие задачи, так и улучшать решение предложенное в данной работе.

Список литературы

[1] Apache Hadoop. — 2015. — URL: http://hadoop.apache.org/ (дата обращения: 21.05.2015).

[2] Apache Pig. — 2015. — URL: http://pig.apache.org/ (дата обращения: 21.05.2015).

[3] Asteriou D. Hall S. ARIMA Models and the Box–Jenkins Methodology // Applied Econometrics (Second ed.). — 2011. — P. 265–286.

[4] B. Efron R. Tibshirani. An Introduction to the Bootstrap. — Chapman Hall/CRC, 1993.

[5] CSV // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Comma-separated_values (дата обращения: 21.05.2015).

[6] D. Borsboom G.J. Mellenbergh J. van Heerden. The Theoretical Status of Latent Variables // Psychological Review. — 2003.

[7] Dean Jeffrey, Ghemawat Sanjay. MapReduce: Simplified Data Processing on Large Clusters // OSDI'04: Sixth Symposium on Operating System Design and Implementation. — 2004.

[8] FOAS. RStudio. — 2015. — URL: http://www.rstudio.com/ (дата обращения: 21.05.2015).

[9] Geoffrey K.F. Tso Kelvin K.W. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks // Energy. — 2007. — Vol. 32, no. 9. — P. 1761–1768.

[10] Humidex Index // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Humidex (дата обращения: 21.05.2015).

[11] JL. Elman. Finding structure in time // Cognitive Sci. — 1990. — Vol. 14. — P. 179–211.

[12] Janert Philipp K. Data Analysis with Open Source Tools. — O'REILLY, 2010. — P. 59–132.

[13] Jeffrey A. Greene Scott C. Brown. The Wisdom Development Scale: Further Validity Investigations // International Journal of Aging And Human Development. — 2003.

[14] Law of Large Numbers // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Law_of_large_numbers (дата обращения: 21.05.2015).

[15] Oracle DBMS. — URL: https://www.oracle.com/database/index.html (дата обращения: 21.05.2015).

[16] Quinlan J. R. Induction of Decision Trees // MACH. LEARN. — 1986. — Vol. 1. — P. 81–106.

[17] R^2 score // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Coefficient_of_determination (дата обращения: 21.05.2015).

[18] Robert Gentleman Ross Ihaka. R. — 2015. — URL: http://www.r-project.org/ (дата обращения: 21.05.2015).

[19] SQL // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/SQL (дата обращения: 21.05.2015).

[20] Siegelmann Hava T., Sontag Eduardo D. On The Computational Power Of Neural Nets // JOURNAL OF COMPUTER AND SYSTEM SCIENCES. — 1995. — Vol. 50, no. 1. — P. 132–150.

[21] Taghrid Samak Christine Morin David Bailey. Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area // Renewable and Sustainable Energy Reviews. — 2007. — Vol. 12. — P. 2040–2065.

[22] Taghrid Samak Christine Morin David Bailey. Energy Consumption Models and Predictions for Large-scale Systems // Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW). — 2013. — Vol. 27. — P. 899–906.

[23] XML // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/xml (дата обращения: 21.05.2015).

[24] XSLX // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Office_Open_XML (дата обращения: 21.05.2015).

[25] Интерполяция // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Interpolation (дата обращения: 21.05.2015).

[26] Линейная регрессия // Wikipedia, free encyclopaedia. — 2015. — URL: http://en.wikipedia.org/wiki/Linear_regression (дата обращения: 21.05.2015).