Нетология

Дополнительная профессиональная программа “Аналитика данных” (DA - 26)

Полякова Ольга Леонидовна

Исследование структуры энергетики (поиск инсайтов, составление рекомендаций стейкхолдерам, построение предиктивной модели цен)

Дипломная работа

2022

Содержание работы

Методологическое введение в аналитическое исследование

Развитие энергетики сегодня неразрывно связано с климатической повесткой, поэтому изучение изменения структуры источников электроэнергетики в контексте национальной энергетической отрасли является задачей востребованной и значимой для широкого круга стейкхолдеров. Международные экологические организации, осуществляющие мониторинг экологических показателей атмосферы, правительственные учреждения, призванные претворять в действительность политические решения международных климатических конгрессов, энергопроизводители, стремящиеся как минимум соблюсти свои позиции при неизменном конфликте интересов, порожденным любыми экономическими инновациями, простые обыватели - конечные потребители любого производимого продукта, от повседневного бытового поведения которых зависит в итоге экологическая ситуация на планете - вот, вероятно, неполный список заинтересованных сторон такого исследования. Мера и характер их заинтересованности диктуют основные требования, выдвигаемые для проведения исследования.

Так, и сегодня существует необходимость общественного экологического просвещения с целью выработки потребности к рациональному образу жизни, осознанию значимости индивидуальной заботы о защите окружающей среды. Важность формирования нового верифицируемого знания о влиянии человека на окружающую среду невозможно переоценить. Здесь помимо широкого социального интереса свое участие проявляют и институализированные экологические организации, призванные еще в большей степени усиливать общественный резонанс относительно рисков ухудшения экологической обстановки на планете. Поэтому выявление неких закономерностей между ростом народонаселения, урбанизацией как значительно более энергоемким социальным образом жизни, экономическим ростом и производством энергии, галопирующим ростом выбросов парниковых газов будет способствовать людям, неравнодушно и ответственно воспринимающим свою миссию экологического просвещения, быстрейшего достижения своих целей.

Правительственные органы, уполномоченные в национальном масштабе координировать выполнение страной своих обязательств в рамках новой международной экологической политики, заинтересованы в достоверной информации о том, насколько методично и организованно выполняются те или иные решения регуляторов, в какой мере их предписания воздействуют на энергетические и экологические показатели развития национальной экономики.

Для хозяйствующих субъектов получение надежных действительных данных имеет критическое значение для управления рисками роста издержек производства и упущенной выгоды в контексте трансформации энергетической отрасли. Представляется, что именно предприятия, окажутся главными выгодоприобретателями владения и управления данными о процессах, происходящих в энергетике, как подлинным корпоративным активом.

Заметим, что наша гипотеза не отличается оригинальностью, но мы ожидаем, что одновременно с многочисленными аналогичными исследованиями ее доказательство будет способствовать приближению к долгосрочной цели сохранения экологического

статуса в мире. Мы полагаем, что внедрение data-driven подхода к управлению, учету, анализу энергетических данных приведет внедривших его предприятия к существенному снижение энергоемкости производства, повышению эффективности энергопотребления, повороту к новым источниками энергии.

Как уже отмечалось, исследование делится на две основные части, первая из которых создает глобальный контекст развития национальной энергетики, аналитике которой посвящена вторая часть работы. Справочная информация представлена отдельными, по большей части несогласованными между собой наборами данных о различных социальных, экономических, экологических, энергетических аспектах исторического развития и современного состояния мировой цивилизации, исследование испанской энергетической системы более сфокусировано и направлено на выявление зависимостей потребления энергии от внешних (погодных) факторов, ресурсов для наращивания производства энергии в моменты наивысшего спроса на нее. Особое внимание уделяется созданию прогноза структуры источников энергии и предикативной модели будущей цены на электроэнергию.

Сегодня аналитическому изучению и обработке данных о мировой и национальных энергетиках посвящено значительное число исследований, информационных ресурсов, баз данных. Google-запрос выдает результат в более, чем 8 миллионов страниц. Наиболее объемные источники данных содержатся в базе данных энергетической статистики Статистического отдела ООН [UNdata], ежегодном Статистическом обзоре мировой энергетики BP [Statistical Review of World Energy | Energy economics | Home], информационном портале [Our World in Data]. В данной работе предпринята лишь учебная попытка индивидуального преобразования и анализа этого обширного материала в глобальном и национальном контексте. Особенностью исследования стала разработка предиктивной модели цены и прогноза будущей структуры источников электроэнергии в Испании.

Аналитическая справка о глобальных экологических и энергетических процессов

Энергетика - та отрасль мировой экономики, в которой учету и сбору данных на каждом этапе ее функционирования традиционно уделяется огромное внимание. Энергоэффективность стала краеугольным камнем развития отрасли, поскольку вся энергетическая система прежде всего подчиняется фундаментальному физическому закону энтропии. Энергия, необходимая для тепла, света и работы машин, не может складироваться в виде сырья или готовой продукции, напротив, она рассеивается, но таким образом, она не только становится бесполезной, но и обогревает атмосферу. И если в индустриальную эпоху промышленные предприятия должны были постоянно подчиняться целям сохранения и экономии энергии, то в постиндустриальную эру на первый план вышла проблема охраны окружающей среды вследствие зачастую разрушительного воздействия на нее со стороны постоянно увеличивающих потребление энергии производств, сфер управления и обслуживания и домохозяйств.

Смена глобальной энергетической парадигмы сегодня диктует уже не только необходимость организации учета энергопроизводства и энергопотребления, но и

востребует точное прогнозирование энергетического рынка. К примеру, сетевая электроэнергетика всегда являвшаяся объектом строгого учета, на нынешнем этапе развития отрасли переживает настоящую “цифровую революцию”, когда помимо внедрения системы “интеллектуального” учета и индивидуального подхода к потребителям и оптимизации затрат на передачу электроэнергии, в энергетику пришли информационные технологии обработки больших данных и сквозной аналитики. Благодаря им data driven ориентированные компании добились огромных преимуществ в анализе и прогнозе энергоэффективности и увеличении нормы прибыли бизнеса [См.: Почему большие данные — будущее энергетики :: РБК Pro].

Данная работа - это учебное исследование, направленное на определение возможностей прогнозирования в национальной энергетике на основе набора данных об энергетическом секторе Испании как области применения машинного обучения. Датасет “Hourly energy demand generation and weather. Electrical demand, generation by type, prices and weather in Spain” любезно предоставлен Nicholas Jhana и размещен на платформе Kaggle [См.: Hourly energy demand generation and weather. # Electrical demand, generation by type, prices and weather in Spain]. Набор данных содержит почасовые данные о производстве электроэнергии из различных источников в Испании в 2015-2018 гг, ее цене и потреблении, полученные с общедоступного портала данных службы передачи энергии (ENTSOE) и Red Electric España. Датасет был дополнен данными о погоде от Open Weather API для 5 крупнейших городов Испании.

Для исследования на основе имеющихся данных были сформулированы следующие задачи:

- обозначить наиболее существенные глобальные тенденции, сопровождающие современную энергетическую трансформацию;
- охарактеризовать существующую структуру источников энергии в Испании в 2015-2018 гг.;
- проанализировать зависимость спроса, производства и цены электроэнергии от внешних факторов - погоды;
- определить наиболее гибкие к изменению спроса источники энергии в Испании;
- спрогнозировать цену на электроэнергию;
- смоделировать будущую структуру энергетики по источникам энергии, опираясь на существующие тенденции.

В результате решения первой задачи была получена аналитическая справка об актуальном состоянии мировой энергетической системы, важнейших причинах, приведших к такому состоянию, и ожидаемых следствиях. Анализ проводился на основе данных, полученных с портала Our World in Data.

Анализируя исторические тенденции развития мировой энергетики, прежде всего приходится отметить глобальность и универсальность ее значения в развитии цивилизации в сочетании с насущной необходимостью ее для жизни каждого отдельного человека. Энергия в различных своих формах, ее производство и трансформация для жизни людей сопровождает всю человеческую историю. Но если в доиндустриальную и раннюю индустриальную эпоху домовладелец был вынужден ежедневно индивидуально заботиться о поддержании тепла в своем жилище и приготовлении пищи, обеспечивая при этом своим трудом и энергией создание

необходимых для жизни вещей и продуктов, то современный городской человек (а их сегодня около 56% в мире, тогда как в 1960 году было еще только 33%) [см. *диаграмма 1*], как правило, живет внутри централизованно обогреваемого пространства, потребляя пищевые и не только продукты, полученные из первичного сырья, подверженного промышленной переработке с использованием многократно трансформированной энергии. Ценой такой благоустроенной жизни современных людей становится тотальное увеличение производства и потребления энергии [см. *диаграмма 2*], чрезвычайно драматически отразившееся на экологической ситуации и будущем человечества [всего несколько диаграмм из множества подобных вполне способны убедить в серьезности этих опасений относительно глобального будущего если и не для всей планеты (хотя этот вариант более, чем вероятен), то для человеческой цивилизации: См.: *диаграмма 3* - глобальное потепление, *диаграмма 4* - концентрация парниковых газов в атмосфере].

Безусловно, мировое сообщество уже довольно длительное время назад обратило внимание на глобальное изменение климата, вызванное в частности выбросами парниковых газов: принятая в 1992 году Рамочная конвенция ООН об изменении климата и Киотский протокол 1997 года обозначили первую цель международных усилий - стабилизировать уровень концентрации парниковых газов в атмосфере на уровне, который не допускал бы опасного антропогенного воздействия на климатическую систему планеты. Пришедшее в 2015 году на смену истекшему Киотскому протоколу Парижское соглашение по климату сформировало еще более строгие цели и ограничения - максимально быстро достичь пика эмиссии углекислого газа, удержать рост глобальной температуры в пределах 1,5℃. Основными путями реализации этих целей являются снижение выбросов и достижение чистых нулевых выбросов парниковых газов - углеродной нейтральности к 2030-2050 годам, декарбонизация экономики в целом, т.е. переход к производствам, основанным на низкоуглеродных источниках энергии. [См. *диаграмму 5* - Корреляция между годовым изменением производства низкоуглеродной энергии и возобновляемой энергии и годового изменением потребления ископаемого топлива].

Особенностью Парижского соглашения стало отсутствие какого-либо механизма принуждения в достижении самостоятельно установленных национальных целей по сохранению климата. Осознание собственной ответственности и рациональная линия поведения в отношении природы - теперь это не только исключительная обязанность политических властей или прерогатива крупных хозяйствующих субъектов, но и реальная забота каждого человека.

Однако зачастую справедливый пафос экологически обеспокоенных активистов упирается в другое ограничение - значительный дефицит ресурсов для проведения преобразований. Экономическая эффективность предпринимаемых технических инноваций оказывается абсолютно не праздным вопросом для многих участников политического процесса сохранения климата.

Кроме вышеуказанного дисбаланса эксперты по глобальным экологическим проблемам отмечают появление новых рисков в решении климатической повестки, как геополитических - чреватых усилением конкурентной борьбы против политических оппонентов с целью доминирования на мировом рынке энергоносителей [См.: Декарбонизация как инструмент конкурентной борьбы против российских компаний -

Ведомости], так и опасности нового роста парниковых эмиссий, компенсируемых технологиями их улавливания [См.: Коммерсантъ - Чистые нулевые выбросы нечисты]. Вопрос о декарбонизации как инструменте климатического регулирования импорта и непрозрачность понятия и технологий “чистых нулевых выбросов” потребуют еще очень долгого и взвешенного экспертного анализа, а учитывая, что потенциально возможное его решение получает абсолютную значимость в самом глобальном контексте, весь процесс нахождения этих решений будет сопровождаться сбором, накоплением и анализом больших данных - Big data. И достижение целей реальной экологической трансформации в энергетике будет возможно только при применении data driven подхода.

Анализ и прогноз энергетических показателей Испании на основе data-driven подхода

В Европейском союзе новая экологическая политика является основной темой современной повестки. ЕС планирует серьезное сокращение выбросов парниковых газов в атмосферу и резкое сокращение использования ископаемых видов топлива для производства энергии. С 2023 года планируется введение программы трансграничного углеродного регулирования, что среди прочего предполагает введение новых налогов на импорт товаров и продукции из стран с высокими выбросами CO2, метана и других вредных веществ в атмосферу Земли.

Испания, являясь с 1985 года полноправным членом Евросоюза и постоянно ощущая свою принадлежность к ядру ЕС, нацелена на реализацию долгосрочной европейской экологической политики. Поэтому все принятые на себя обязательства в области сохранения климата всей единой Европой касаются и национальных целей развития испанской экономики.

Что касается энергетического комплекса в структуре испанской экономики, то он является функцией энергоемкости ее производства и обслуживания, с одной стороны, и климата, с другой стороны. Несмотря на то, что страна практически постоянно испытывает системные трудности экономического развития и нестабильный рост ВВП, она является одним из лидеров в Европе в области высокотехнологичного производства, связанного, как правило, с рациональным подходом в сфере энергоэффективности [См.: Экономика Испании — Википедия]. (Однако изучение этого вопроса находится за рамками нашего исследования.)

Одной из задач настоящего исследования является анализ того, как погода в качестве внешнего фактора влияет на динамику спроса и предложения электроэнергии в национальном масштабе. Естественно, что климат определяется географическим местоположением. И Испания в этом отношении территориально достаточно крупное государство, расположенное на Пиренейском полуострове, омываемом водами Средиземного моря и Атлантического океана. Испания - южно-европейское государство, ее климат характеризуется высокой среднегодовой температурой в 16,48 ℃ и большим количеством солнечных дней в году (83% времени) [См.: расчет показателей среднегодовой температуры и количества малооблачных и солнечных дней в году].

Наш анализ зависимости уровня энергопотребления от температурных изменений не дал положительных результатов. Вероятно, для обнаружения связи между показателями погоды и спроса на электроэнергию необходимо включать большее количество факторов. В целом можно констатировать, что динамика потребления электроэнергии в национальном масштабе носит более сложный характер, и ее сезонность определяется причинами, лежащими за рамками доступных нам данных.

Возможность определения многофакторной зависимости мы находим в решении вопроса о влиянии различных видов генерации энергии на общую генерацию с целью удовлетворения спроса на электроэнергию (общее потребление). Проведенный анализ показал, что пиковые нагрузки покрываются за счет увеличивающегося производства энергии из ископаемого газа. Этот источник можно смело отнести к наиболее гибким видам электрогенерации.

Анализ ситуации на энергетическом рынке Испании, сложившейся в 2015-2018 годах, на основе имеющихся данных можно завершить определением структуры источников производства электроэнергии. В контексте актуальных трендов всю электрогенерацию разделяем на группы высокоуглеродных и низкоуглеродных источников энергии, сопоставляя объемы производства в целом для этих групп друг с другом. Что касается высокоуглеродных источников энергии, то следует отметить, что Испания практически не обеспечивает себя углеводородными источниками самостоятельно, традиционно являясь их реципиентом на внешнем рынке [См.: диаграмму 8. Источники покрытия пиковых нагрузок].

До поворота к тотальной декарбонизации энергетики недостаток собственных углеводородов компенсировался в Испании развитием атомной энергетики, кроме этого географическое положение страны позволяет ей широко использовать возобновляемые источники энергии - ветер и солнце, энергию гидроэлектростанций. Однако можно ли говорить о том, что в Испании действительно в период с 2015 года - года подписания Парижского соглашения о климате - получил развитие поворот к декарбонизации экономики, т.е. в относительном выражении низкоуглеродная энергетика приобрела в сравнении с энергией из ископаемого топлива преобладающее значение, или и по сей день сохраняются прошлые тренды развития энергетики.

Простая визуализация соотношения высокоуглеродных и низкоуглеродных источников в испанской энергетике свидетельствует о преобладании генерации низкоуглеродной энергии в 2015-2018 гг. [См. *диаграмму 9.* Доли высокоуглеродных и низкоуглеродных источников энергии в общей генерации в 2015-2018 гг.] Но в то же время, определив циклический характер изменений структуры энергии, мы не можем судить о направленности динамики в использовании той или другой группы источников энергии в будущем. Чтобы сделать аргументированное предположение о дальнейшем развитии энергетической отрасли, необходимо найти иной способ прогнозирования. При выборе нового подхода к изучению данной проблемы мы отталкивались от факта того, что форма представления данных в наших датасетах - это временные ряды, а традиционно сферой в наибольшей степени вбирающей в себя множество методов анализа динамических временных рядов является биржевой теханализ, из которого и был почерпнут индикатор Аллигатор, разработанный Б.Вильямсом. Индикатор Аллигатор обычно в теханализе используется для определения направления тренда и,

поэтому в некоторой степени может предупреждать о дальнейшем развитии процесса. Прибегнув к авторской (Б.Вильямса) интерпретации индикатора в отношении исследуемых данных, на различных (месячном и дневном) временных периодах мы в данном случае не можем прийти к вполне определенным выводам об ожидаемых изменениях в динамике [См.: Аллигатор: как «хищный» индикатор может помочь трейдеру?]. Направленность наших ожиданий будущего состояния энергетических процессов зависит от рассматриваемого временного периода. Но если в теханализе эта ситуация вполне допустима в силу фрактального характера движения финансовых показателей, то в сфере проектирования и анализа производственных процессов необходимы гораздо более точные и определенные прогнозы [См. *диаграммы 10-11*] .

Выход из данной противоречивой ситуации был найден в использовании новой библиотеки Python *pmdarima* для анализа временных рядов. В результате применения последовательных этапов *модели ARIMA* и использования функции *auto_arima* указанной библиотеки были сформированы предиктивные модели развития высокоуглеродной и низкоуглеродной энергетики, а также предложен прогноз будущего значения показателей этого развития [См.: *Smith, Taylor G., et al. pmdarima: ARIMA estimators for Python, 2017-*, ARIMA estimators for Python — pmdarima 1.8.4 documentation].

В исследовании также была реализована еще одна задача, имеющая немаловажное значение для разработки стратегий национальной энергетики и экономики в целом - построение предиктивной модели будущей цены на электроэнергию в Испании с помощью алгоритмов *machine learning*.

Справка о применении data-driven подхода в исследовании

Справочная информация в исследовании получена в результате аналитической обработки данных с портала Our World in Data. (Этот портал специализируется на сборе, анализе, визуализации глобальных данных по различным социально-экономическим и гуманитарным проблемам, он предоставляет статистические данные, код обработки и визуализации бесплатно без каких-либо ограничений. В данной работе я с благодарностью к его авторам(Dr. Max Roser, Dr. Hanna Ritchie и др.) воспользовалась csv-файлами с наборами данных по соответствующим аспектам исследования). В целом исходные данные были достаточно качественными и не нуждались в дополнительной очистке.

Для *диаграммы 1. Урбанизация населения в мире* на основе датасета, содержащего данные о количестве городского и сельского населения в странах мира с 1960 по 2020 гг. выделены данные по миру в целом, рассчитаны проценты, и построена популяционная диаграмма, скорректированная на данные об урбанизации.

Для *диаграммы 2. Линейный график мирового потребления энергии* источником послужил датасет о погодовом потреблении энергии в разных странах и мире в целом. Выделив данные о мировом потреблении с 1965 по 2019 гг., строим линейный график.

Для *диаграммы 3. Линейный график глобального изменения температуры и скользящей средней температурной аномалии* данные получены из набора данных о комбинированных аномалиях температуры воздуха у поверхности суши и воды у поверхности моря, рассчитанных как отклонения от среднего значения за 1951-1980 гг. Поскольку последнее значение временного диапазона для определения нулевой отметки температурных аномалий приходится на 1980 г. расчет скользящей средней температурных отклонений мы начинаем с 1981 года. Для значений температурных аномалий за весь период наблюдений и полученных значений скользящей средней с 1981 года строим наложенные друг на друга линейные графики.

Для *диаграммы 4. Сравнительный график концентрации парниковых газов в атмосфере* были загружены данные из трех датасетов, посвященных динамике концентрации газов в атмосфере, способствующих скорейшему наступлению парникового эффекта, - углекислому газу, метану, диоксиду азота. Подготовка данных к визуализации включала приведение данных к одному и тому же периоду и построение трех линейных графиков в одном контейнере.

Для *диаграммы 5. Корреляция между годовым изменением производства низкоуглеродной энергии и возобновляемой энергии и годовым изменением потребления ископаемого топлива* использованы наборы соответствующих данных. Полученные датафреймы приведены к значениям в целом по миру и объединены по столбцу ‘Year’. С помощью метода pandas .corr(method=’spearman’) вычислены коэффициенты корреляции между полученными рядами значений годовых изменений в энергетике и построена визуализация для корреляционной матрицы.

Созданная корреляционная матрица и ее визуализация показывает, что существует достаточно большая корреляция между процессами производства низкоуглеродной энергии и возобновляемой энергии (что и понятно, поскольку их источники во-многом совпадают). Но особенно важно, что обнаружена пусть и небольшая, но отрицательная

корреляция между потреблением ископаемого топлива и производством низкоуглеродной энергии.

Изучение испанского датасета

Для получения информации о климатических особенностях Испании будем использовать датасет с почасовыми данными о температуре и других погодных показателях для 5 крупнейших городов. Предварительный анализ данных показывает отсутствие пропусков и существенных аномалий в данных. Датафрейм готов к работе.

Произведем расчет показателя среднегодовой температуры в Испании. Принцип расчета: группируем по 5 представленным городам, агрегируем данные по среднему, суммируем средние показатели по городам и приводим их к среднему, переводя при этом температурные показатели в кельвинах в градусы Цельсия. Для показателя времени ясного и малооблачного неба принцип расчета заключается в приведении количества часов ясного неба и небольшой облачности к общему времени наблюдений.

Для *диаграммы 6. Существует ли зависимость между погодными условиями и уровнем потребления энергии* и дальнейшей аналитики создаем датафрейм с почасовыми данными генерации электроэнергии из различных источников и общего потребления энергии, дневными прогнозами облачности и ветра, общего спроса и цены электроэнергии. Проверяем и чистим датафрейм, для чего сначала удаляем столбцы с полностью отсутствующими данными. В оставшихся столбцах по-прежнему еще сохраняются пропуски, необходимо принять решение о процедуре восполнения или удаления оставшихся пропусков. Определяя дни с наибольшим количеством пропущенных значений, приходим к выводу, что пропуски - случайны, это позволяет выбрать способ заполнения пропусков предыдущими значениями. Полагаем, что количество строк с пропусками, составляющее 0,13% к общему числу строк, не должно оказать существенного искажающего результат влияния. Подготовив данные к анализу, переходим к установлению характера зависимости общего потребления энергии от погодных условий.

Преобразуем строковые значения записи даты в датафрейме в объект datetime для группировки часовых значений в дневные, месячные и годовые. Для анализа будем использовать дневной таймфрейм. Для оценки динамики потребления энергии построим ее линейный график. График показывает, что колебания дневного потребления энергии происходят в диапазоне 600000 - 800000, в отдельные дни наблюдений превосходя верхнюю границу диапазона или, напротив, опускаясь ниже 600000.

Аналогично данному датафрейму в weather мы также приводим показания к дневным измерениям и выделяем данные по Мадриду. Освобождаясь от избыточных данных, объединяем датафреймы и вычисляем коэффициент корреляции между двумя рядами значений - потребление энергии и средняя дневная температура в Мадриде. Получившееся значение очень четко показывает, что значимой зависимости между температурой и потреблением энергии не существует.

Для *диаграммы 8. Источники покрытия пиковых нагрузок* первоначально определим коэффициенты корреляции общего потребления и различных видов генерации. Числовые ряды значений с наибольшим коэффициентом корреляции выделим из общего объема данных, отфильтруем значения, превышающие пороговое значение для пиковых нагрузок (> 800000) и для этих данных построим диаграмму рассеяния. Из диаграммы становится видно, что пиковые нагрузки в энергопотреблении покрываются за счет генерации из ископаемого газа [См. *диаграмму 8*].

Для *диаграммы 9. Структура источников энергии* на годовом таймфрейме группируем данные по критерию высокоуглеродные/низкоуглеродные источники, создаем круговые диаграммы для каждого с 2015 по 2018 гг. года наблюдений [См.: диаграмму 10. Структура источников энергии]. Построенная диаграмма, однако, не дает какого бы то ни было определенного направления для дальнейшей динамики. Но, как известно, одной из важных задач data-аналитики является прогнозирование последующих показателей процессов.

С этой целью был предложен заимствованный из арсенала биржевого теханализа индикатор Аллигатор, представляющий собой комбинацию трех скользящих средних с различными временными интервалами, которые смещены на определенное количество периодов. В силу того, что Аллигатор способен определять направление тренда, его можно использовать и в качестве предсказывающего индикатора, исходя из предположения, что существующее направление тренда сохранится еще какое-то время [См. *диаграммы 10 и 11. Индикатор Аллигатор Вильямса для Iдневного и месячного графика генерации энергии*].

Совершенно очевидно, что подобного вида прогноз, во-первых, обладает не слишком хорошей надежностью и, во-вторых, довольно плохо предсказывает в ситуации “флэта” - горизонтального изменения показателей, не характеризующегося признаками тренда, кроме того, в нашем исследовании он не показал достаточно определенных результатов при применении на дневном и месячном временном периоде. Неудовлетворительность результата в достижении задачи прогнозирования будущего состояния в энергетике Испании потребовала от нас дальнейшего поиска других возможных решений проблемы.

Основная специфика исследуемого в работе набора данных - то, что данные представлены в виде временного ряда (timeseries), т.е. последовательности в которой метрики записаны через регулярные промежутки времени, в нашем случае - ежечасно. И указанная специфичность проявляется прежде всего в том, что в аналитике и прогнозировании подобных данных их временная привязка получает определяющее значение. Как правило, внутренняя природа timeseries раскрывается через совокупность таких компонентов, как базовый уровень, тренд, сезонность, ошибка, последние три из которых могут оказывать различное влияние на характер изменения метрики. Анализ временного ряда и его подготовка к прогнозированию в любом случае должны включать оценку этих факторов в формирование timeseries. Кроме этого, в прогнозировании будущих значений timeseries можно выделить два основных подхода: прогноз одномерного временного ряда, когда для предсказания используются только его предыдущие значения, и экзогенное прогнозирование в случае, когда в качестве предикторов используются данные, отличные от предсказываемого временного ряда [См.: Time Series Analysis in Python - A Comprehensive Guide with Examples - ML+].

В статистике и Data Science на сегодняшний день разработано уже множество различных алгоритмов прогнозирования, и конкретно для временных рядов наиболее часто рекомендуемым является ARIMA (AutoRegressive Integrated Moving Average), алгоритм, объясняющий временной ряд на основе его собственных прошлых значений: собственных лагов и запаздывающих ошибок прогноза. В классическом варианте для моделирования с помощью алгоритма ARIMA требуется выполнение нескольких условий: временной ряд должен быть освобожден от сезонности и не являться случайным “белым шумом”. К тому же, наилучшие результаты ARIMA показывает на стационарных временных рядах. Поэтапная подготовка данных к работе ARIMA предполагает, что, применяя к ряду функцию автокорреляции (ACF), можно установить наличие сезонности и необходимость дифференцирования членов ряда для преобразования последовательности в стационарный ряд. Эта процедура позволит определить первый параметр модели *d* - количество разностей, необходимых для того, чтобы ряд стал стационарным [См.: *диаграммы 12 и 14. Изучение данных для построения модели ARIMA*].

Другими необходимыми параметрами модели ARIMA являются *p* - порядок авторегрессионной модели AR - иначе количество временных лагов прогнозируемой величины, которые будут использоваться в качестве предикторов, q - порядок модели скользящего среднего (MA), или величина “окна” данных, используемого в расчете. Важнейшая цель подготовительного этапа - определить значения p, d, q, а также s - если временной ряд обладает сезонностью, т.е. повторяемостью динамики значений ряда через определенные промежутки времени, и использовать при построении модели ARIMA.

Однако в последнее время были разработаны автоматизированные алгоритмы прогнозирования, существенно облегчающие работу с данными. В новой python-библиотеке pmdarima ее вдохновитель и автор Тейлор Смит расширил прогностические возможности Python до уровня некогда превосходившего его в этом аспекте R, спроектировав новую модель *auto_arima* для Python как можно более близкой к аналогичной R-модели. При этом важнейшим достижением становится объединение всех моделей ARIMA различных библиотек статистического и машинного обучения (statsmodels и scikit-learn) в один класс.

Алгоритм *auto_arima* выбирает лучшую ARIMA-модель для одномерного временного ряда в соответствии с минимальным AIC или другим критерием для выбора модели прогнозирования, основанным на оценке объема потерянной в результате работы алгоритма информации.

Применение *auto_arima* имеет преимущество в том, что модель автоматически определяет наиболее оптимальные параметры, рационально подбирает их правильную комбинацию, внутри себя проводя испытания с различными параметрами, заключенными в диапазоны. Модели временных рядов основаны на эндогенной темпоральности, поэтому мы не можем просто разделить данные случайным образом и включаем в тестовую выборку отдельный временной период (25% всего объема данных). Создаем модель, и моделируем процесс на основе наших данных[См.: *диаграммы 13.* Р*еальные значения производства высокоуглеродной энергии и результаты предиктивной модели*].

Поскольку нам необходим прогноз двух процессов, поэтому повторяем всю процедуру с новыми данными - для низкоуглеродной генерации [См.: *диаграммы 14.* Р*еальные значения производства низкоуглеродной энергии и результаты предиктивной модели*].

В плане метрик качества модель *auto_arima* использует MSE, SMAPE, значения которых требуют проведения процедуры подгонки параметров модели. Библиотека pmdarima предлагает пользователям самостоятельно разрабатывать наиболее чувствительные к специфическим данным модели в рамках эстиматора ARIMA. Его применение в нашем исследовании носило экспериментальный характер, но позволило сгенерировать отнесенные в будущее значения показателях. При условии того, что параметры модели будут очень тонко настроены на получение валидных результатов (это, естественно, предмет уже нового исследования), можно говорить о том, что нами была найдена рабочая модель для прогнозирования временных рядов.

Разработка предиктивной модели цены электроэнергии

В исследовании были созданы и протестированы три прогнозных модели машинного обучения: модель линейной регрессии, “случайный лес” и XGBoost, в которые подавались специально подготовленные данные из датафреймов energy_spain и weather. Мы исходили из посылки, что рыночная цена на электроэнергию формируется соотношением спроса и предложения, а также находится под влиянием ожиданий будущего этого соотношения у продавцов и потребителей. Следовательно, нам необходимо было суммировать все виды генерации электроэнергии, чтобы получить величину общего предложения на рынке, включить в совокупность признаков показатели прогнозов солнца и ветра на суше на день вперед, прогноза нагрузки и фактической нагрузки. Для включения внешних факторов изменения цены мы произвели слияние датафрейма почасовой генерации и потребления энергии и выделенного датафрейма часовых погодных данных в Мадриде. Целевой переменной регрессионной модели была определена цена энергии.

В *модели линейной регрессии [LinearRegression]* - одной из классических линейных моделей и по сегодняшний день широко применяемой в статистических исследованиях и машинном обучении в силу простоты и ясности понимания того, как входные данные влияют на результаты моделирования, данные были подвергнуты случайному разделению на обучающую и тестовую выборки в пропорции 0.75/0.25. Однако, несмотря на указанную прозрачность моделирования, оценка линейной регрессии по методу R**2 продемонстрировала ее невысокое качество.

Следующие две модели относятся к ансамблевому обучению, идея которого заключается в “построении модели прогнозирования путем объединения сильных сторон набора более простых базовых моделей”, из которых она и состоит [Хасти, Тр., с. 635]. *RandomForestRegressor* основывается на методе усреднения базовых независимых оценок, приводящем к уменьшению дисперсии. *Модель XGBoost* стремится к уменьшению смещения оценки, скомбинированной из последовательности базовых оценок прогноза. Обе модели принимают сразу весь объем данных для моделирования и выдают в свою очередь довольно близкие высокие оценки.

Если мы продолжим сравнение моделей с целью выбора наиболее предпочтительной из них, то и метрика RMSE подтвердит большую ошибочность прогноза методом линейной регрессии над *RandomForestRegressor* и *XGBoost.* Следовательно, обе эти модели мы можем рекомендовать для использования в прогнозировании будущей цены электроэнергии.

Итоги исследования

Аналитическое исследования испанской энергетики в контексте глобальных экологических и экономико-технологических трендов показало, что реальное достижение декларативных политических целей в вопросах энергоэффективности и сохранения климата - это долгий, сложный, планомерный, требующий постоянного внимания со стороны курирующих экологию международных организаций, национальных правительств, энергетических корпораций, частных потребителей процесс. В сугубо аналитическом плане в работе были решены поставленные задачи и достигнуты верифицируемые результаты на основе развернутой системы метрик. Во-первых, обнаружено, что потребление электроэнергии в Испании в существенной степени не зависит от такого внешнего фактора, как погода. Во-вторых, установлено, что основным ресурсом для покрытия пикового увеличения нагрузки является природный газ. Поэтому на данном технологическом этапе развития полного отказа от использования голубого топлива как наиболее экологичного источника энергии ожидать не приходится. В-третьих, соотношение долей в структуре источников энергии, сгруппированных по критерию содержания в них углеводородов, имеет только циклическое развитие, а разработка иных индикаторов этих показателей также проявляет горизонтальное движение в постоянном диапазоне. Также в исследовании была выявлена возможность числовых прогнозов энергетических показателей, но только при условии тщательной подгонки прогнозной модели. В-четвертых, в работе была представлена предикативная модель будущей цены на электроэнергию в Испании.

Представленные результаты работы позволяют дать достаточно развернутую характеристику состояния национальной энергосистемы, но также содержат в себе возможность дальнейшего исследования путем введения новых метрик или более детальной разработки предложенных.