УДК 004.65

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В РАМКАХ ЕДИНОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ГОСУДАРСТВЕННОЙ СТАТИСТИКИ

А.А. Коробчук
Главный специалист Главного статистического управления Брестской области

Главное статистическое управление Брестской области.
E-mail: alinakorobchuk94@mail.ru.

А. А. Коробчук

Окончила Академию управления при Президенте Республики Беларусь по специальности «Управление информационными ресурсами», а также магистратуру по специальности «Государственное управление», профилизация «Электронное правительство». Работает главным специалистом в ГСУ Брестской области. Проводит научные исследования в области интеллектуального анализа статистических данных.

Аннотация. Статистическая информация – одна из основных составляющих государственного информационного ресурса. Сегодня к ее качеству предъявляются всё более высокие требования. Использование технологий интеллектуального анализа данных и машинного обучения статистическими органами в рамках Единой информационной системы государственной статистки ведёт к повышению качества официальной статистики и созданию инновационной. Рассмотрены способы применения этих технологий статистическими органами зарубежных стран. Проведен анализ данных статистики промышленности и построены прогнозные модели при помощи различных методов анализа временных рядов на языках программирования Python и R, оценено качество построенных моделей.

Ключевые слова: Статистические данные, интеллектуальный анализ данных, машинное обучение, экспериментальная статистика, методы анализа временных рядов, прогнозирование.

Введение.

В последние десятилетия наблюдается повышенное внимание к такому явлению, как «большие данные» (далее – Big Data). Это связано с тем, что объём собираемой, хранимой и передаваемой информации стремительно увеличивается. Увеличивается и объём статистической информации, которая является одной из основных составляющих государственного информационного ресурса. На основе нее принимаются важные политические решения, она предоставляет адекватную оценку ключевых экономических и социальных показателей, охватывающих все аспекты развития государства.

На сегодняшний день общество и экономика предъявляют все более высокие требования к качеству официальной статистической информации. В связи с этим возникает необходимость повышения качества статистических данных и совершенствования системы статистического производства с помощью использования новых технологий, таких как интеллектуальный анализ данных Data Mining (далее – DM) и машинное обучение Machine Learning (далее – ML).

Технологии интеллектуального анализа больших статистических данных.

Технология Data Mining позволяет извлечь полезные знания, способствуя совершенствованию бизнесстратегий, баз знаний, научных и медицинских исследований. Можно сказать, что интеллектуальный анализ данных – это процесс определения новых, корректных и

потенциально полезных знаний на основе больших массивов данных. Также Data Mining можно определить как выявление скрытых закономерностей или взаимосвязей в больших массивах необработанной информации [1, с. 5].

Теория обучения машин находится на стыке прикладной статистики, численных методов оптимизации, дискретного анализа и применяется с середины 20-го века, в то время, как интеллектуальный анализ данных более молодая дисциплина, необходимость в ней возникла в конце 20-го века. По составу решаемых задач Data Mining практически не отличается от стандартного набора средств в области Machine Learning. Основное различие заключается в эффективности алгоритмов и технологичности их применения [2].

Среди методов машинного обучения выделяют логические методы, линейные методы, метрические методы и другие. Существуют также композиционные методы в машинном обучении – случайный лес и градиентный бустинг. Композиции линейных моделей классификации и регрессии представляют отдельный класс методов – нейронные сети [3].

Поскольку статистические данные представляют собой данные в виде временных рядов различных показателей, следует отдельно отметить методы, которые используются при анализе временных рядов. Временной ряд – совокупность значений какого-либо показателя или процесса, собранная в различные моменты времени. Задачи, решаемые при анализе временных рядов, можно разделить на две большие группы.

1. Анализ временных рядов.
2. Прогнозирование временных рядов.

Для решения первой задачи применяются методы выделения тренда, поиска сезонных компонент, разложения рядов, поиска аномалий и др. Для решения второй задачи применяются авторегрессионные модели (AR), скользящее среднее (MA), ARMA, ARIMA, адаптивные модели и другие [4].

Применение методов интеллектуального анализа данных в органах государственной статистики в зарубежных странах.

На сегодняшний день статистические органы зарубежных стран уже используют методы DM и ML при производстве как официальной, так и инновационной статистики. В ходе анализа способов применения данных методов органами статистики передовых зарубежных стран было выявлено, что Статистические службы Европейского союза (Евростат), Эстонии, Нидерландов, Швейцарии и других стран занимаются разработками Экспериментальной статистики (Expiremental Statistics).

Экспериментальная статистика предполагает производство инновационной статистики с использованием новых источников данных, новых методов ML, DM, а именно методов кластеризации, классификации, регрессии, анализа временных рядов, нейронных сетей и других [5].

Список статистических служб Европейского Союза, занимающихся экспериментальной статистикой.

1. Евростат.
2. Дания, Статистическое управление.
3. Германия, Федеральное статистическое управление.
4. Испания, Национальный институт статистики.
5. Италия, Национальный институт статистики.
6. Латвия, Центральное статистическое бюро.
7. Нидерланды, Центральное статистическое управление.
8. Польша, Центральное статистическое управление.
9. Португалия, Национальный статистический институт.
10. Румыния, Национальный институт статистики.
11. Финляндия, Статистическое управление.
12. Исландия, Статистическое управление.
13. Швейцария, Федеральное статистическое управление.

Экспериментальная статистика нацелена на повышение качества официальной статистики (увеличение скорости обработки, точности, уровня детализации), но и также на создание новой статистики (инновационные проекты, продукты, приложения, услуги). Целями экспериментальной статистики также являются глубокий анализ статистических данных и построение прогнозов.

В статистическом управлении Нидерландов функционирует Центр статистики больших данных (Center for Big Data Statistics, далее – CBDS). Используя методы интеллектуального анализа данных и искусственного интеллекта, CBDS занимается разработкой инновационных продуктов и исследований. Целью функционирования центра является повышение скорости обработки, точности качества официальной статистики, а также создание новых проектов. Среди инновационных продуктов центра – «Определение типов инноваций компаний с помощью анализа текстов» на веб-сайтах компаний; «Оценка количества грузовых автомобилей на дорогах без датчиков», проект используется для определения интенсивности движения грузовиков в транспортной сети; «Измерение цен на коммерческую недвижимость» в сотрудничестве с другими организациями, проект является новаторским из-за широкого охвата данных [6]. Помимо этого на сайте Центрального статистического управления Нидерландов (Statistics Netherlands, далее – CBS) присутствует информация о том, что орган предлагает проведение индивидуальных исследований по запросу на платной основе [7].

Федеральное статистическое управление Швейцарии (Bundesamt für Statistik, далее – BFS) в соответствие с инновационной стратегией в области данных от 21 ноября 2017 года стремится к использованию дополнительных методов анализа, прогнозного анализа с использованием передовых статистических методов, машинного обучения, интеллектуального анализа данных. Пять пилотных проектов были запущены в рамках стратегии инноваций в области данных и в настоящее время реализуются. Один из проектов «Проверка достоверности с машинным обучением» направлен на расширение и ускорение проверки достоверности статистических данных в BFS с использованием алгоритмов машинного обучения, что ведет к повышению качества данных. Этот подход основывается на алгоритме, использующем исторические данные. На основе предыдущего анализа данных определяется целевая переменная, которая должна быть предсказана алгоритмом. Только после этого алгоритм используется для предсказания. На заключительном этапе прогнозируемые и фактические значения целевых переменных сравниваются, и оценивается точность прогнозов. Также используется механизм обратной связи для отправки автоматического объяснения респондентам, предоставляющим данные.

Ещё один проект «Автоматизация кодирования экономической деятельности компаний с использованием методов машинного обучения», его целью является улучшение качества кодирования компаний. Качество кодирования предприятий, зарегистрированных в Регистре предприятий, напрямую влияет на результаты структурной, экономической статистики, касающейся предприятий. Проект направлен на автоматизацию присвоения предприятиям кодов экономической деятельности [8].

В Эстонии среди завершенных проектов – приложение об оплате труда в Эстонии, предоставляет углубленную статистику в графиках и диаграммах по различным профессиям, территории, полу и т. д. Из других проектов: «Анализ социальных групп», «Обзор маленьких регионов и групп населения».

В Евростате также разработано множество проектов на основе большого охвата данных и новых методов DM и ML. Все проекты разбиты по категориям – цены на продовольственные товары; доходы, потребление и благосостояние; торговля услугами и т. д.

В Управлении статистикой Англии функционирует Data Science Campus, который был создан для изучения использования новых источников Big Data и новых методов обработки данных к статистическим данным, уже несколько проектов было запущено [9].

Таким образом, методы машинного обучения и интеллектуального анализа данных уже используются статистическими органами некоторых передовых стран, с помощью них разрабатываются инновационные проекты и инструменты.

Применение методов анализа временных рядов к данным Национального статистического комитета.

Применение методов DM и ML органами статистики ведет к повышению качества официальной статистической информации и предоставлению новой, созданию инновационных продуктов и оказанию новых информационных услуг.

Применять эти методы необходимо в рамках Единой иной информационной системы государственной статистики (далее – ЕИСГС). ЕИСГС – информационная система, разработанная в рамках реализации Государственной программы создания ЕИСГС Республики Беларусь, утвержденной Указом Президента Республики Беларусь от 13 ноября 2006 г № 665. В состав подсистем ЕИСГС входят следующие подсистемы:

- подсистема ведения базы метаданных;
- подсистема сбора и обработки первичных статистических данных;
- подсистема накопления и хранения первичных статистических данных и статистической информации;
- подсистема анализа и распространения статистической информации;
- подсистема обеспечения информационной безопасности.

В ходе анализа подсистем ЕИСГС, было выявлено, что на сегодняшний день методы интеллектуального анализа данных и машинного обучения не применяются Национальным статистическим комитетом. Предлагается внедрить подсистему интеллектуального анализа статистических данных в структуре ЕИСГС. Эта система позволит выявлять скрытые закономерности, новые полезные знания, получать прогнозирующие результаты, необходимые для принятия решений государственными органами, коммерческими и некоммерческими организациями с целью повышения эффективности их функционирования.

Примером информационной услуги с использованием методов DM и ML может служить услуга «Проведение индивидуальных исследований на основе статистических данных», которая может предоставляться по запросу от коммерческих, некоммерческих, а также государственных организаций.

Одним из вариантов исследований может быть построение прогнозов на основе статистических данных. Как отмечалось выше, статистические данные представляют собой данные в виде временных рядов. Все данные Национального статистического комитета можно разделить по следующим направлениям [10].

1. Демографическая и национальная статистика.
2. Экономическая статистика.
3. Многоотраслевая статистика.

Для решения поставленной задачи были получены данные по статистике промышленности (экономическая статистика) по запросу в Национальном статистическом комитете. Данные содержат информацию об объёмах производства и запасов отдельных видов продукции, начиная с января 2016 года по апрель 2020 года, крупных и средних промышленных предприятий мебельной отрасли Брестской области. Всего в полученных данных представлено 42 кода промышленной продукции. Другой набор содержит данные, которые являются первичными обезличенными данными. Согласно Закону «О государственной статистике», первичные данные являются конфиденциальными, однако в обезличенном виде могут быть получены по запросу в органах государственной статистике для определенных целей [11]. Полученные данные содержат информацию об объёмах производства и запасов отдельных видов продукции некоторых промышленных предприятий, начиная с января 2016 года по апрель 2020 года. Всего в полученных данных представлено 38 кодов продукции и 273 промышленных предприятия Республики Беларусь, которые невозможно идентифицировать.

Можно использовать, как аналитические платформы, так и языки программирования для решения поставленных задач. На сегодняшний день существует множество программных решений для анализа данных и машинного обучения, в том числе для анализа временных рядов.

Разнообразные аналитические платформы и фреймворки позволяют проводить продвинутый анализ Big Data. К таким платформам относятся: SAS, STATISTICA Data Miner, Deductor, IBM SPSS Modeler, RapidMiner, KNIME и другие. Однако среди всех существующих инструментов, необходимо выделить языки программирования для интеллектуального анализа Big Data Python и R. R и Python являются незаменимыми инструментами для исследователей данных, также это популярные языки программирования для работы со статистикой, они позволяют написать код четко под нужды разработчика.

Для решения задачи прогнозирования объёма продаж крупных и средних предприятий мебельной отрасли Брестской области на языке Python были использованы эконометрический подход ARIMA, и выпущенная в 2017 году библиотека fbprophet, а затем было оценено качество построенных моделей.

ARIMA – это аббревиатура от AutoRegressive Integrated Moving Average. Модели типа ARIMA – это обобщение модели класса ARMA. Модель ARIMA (p, d, q) – это модель ARMA (p, q) для d раз продифференцированного ряда. Если взять авторегрессионную модель порядка p (AR (p))и модель скользящего среднего порядка q (MA (q))и сложить то, что находится у них в правых частях. Результат – это модель ARMA (p, q), она выглядит следующим образом [12]:

$$y_t=\alpha+\varphi_1 y_{t-1}+\varphi_2 y_{t-2}+\ldots+\varphi_p y_{t-p}+\epsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}\ldots.. \quad (1)$$

Второй способ прогнозирования – использование пакета FProphet. Пакет FProphet был разработан для прогнозирования большого числа различных бизнес-показателей.

По сути, Prophet – это additive regression model, состоящая из следующих компонент:

$$y(t)=g(t)+s(t)+h(t)+t \quad (2)$$

где s (t) – сезонные компоненты, g (t) – тренд, h (t) – компонента, отвечающая за заданные пользователем аномальные дни, t – ошибка.

Качество моделей было оценено при помощи средней абсолютной ошибки MAPE и абсолютной ошибки MAE.

MAPE (mean absolute percentage error) – это средняя абсолютная ошибка прогноза. Пусть yi – это показатель, а $\widehat{y_i}$ – это соответствущий этой величине прогноз модели. Тогда $e_i = y_i - \widehat{y_i}$ – это ошибка прогноза, а $p_i = \frac{e_i}{y_i}$ – это относительная ошибка прогноза.

$$MAPE=mean(|p_i|)\ldots.. \quad (3)$$

Кроме того, бывает полезно смотреть и на абсолютную ошибку MAE – mean absolute error, чтобы понимать, насколько ошибается модель в абсолютных величинах [13].

$$MAE=mean(|e_i|)\ldots.. \quad (4)$$

График общего объема продаж, построенный при помощи библиотеки Matplotlib, на основе полученных данных, представлен на рисунке 1. По оси абсцисс отложен временной ряд, по оси ординат объем продаж в тыс. долл. США.

По графику видно, что после кризиса 2015-2016 года идет плавное увеличение объема продаж, не учитывая сезонные колебания, а затем в марте-апреле 2020 года происходит резкое снижение объема продаж, связанное с непредвиденным для экономики фактором (пандемия коронавирусной инфекции COVID-19).

В ходе анализа данных, была осуществлена STL декомпозиция временного ряда при помощи библиотеки Statsmodels, которая позволяет визуально посмотреть, из каких компонент состоит временной ряд (рисунок 2).

Верхний график – это исходный ряд, следующие, соответственно – тренд, сезонность,

остатки. Тренд увеличивается, сезонность выражена, плавно повышается, доходит до некоторого пика, затем резко снижается. Здесь же проверили критерием Дики-Фуллера стационарен ли временной ряд (рисунок 2.).

```
plt.figure(figsize=(16,8))
plt.title('Total Sales of the industry companies')
plt.xlabel('Time')
plt.ylabel('Sales')
plt.plot(ts);
```

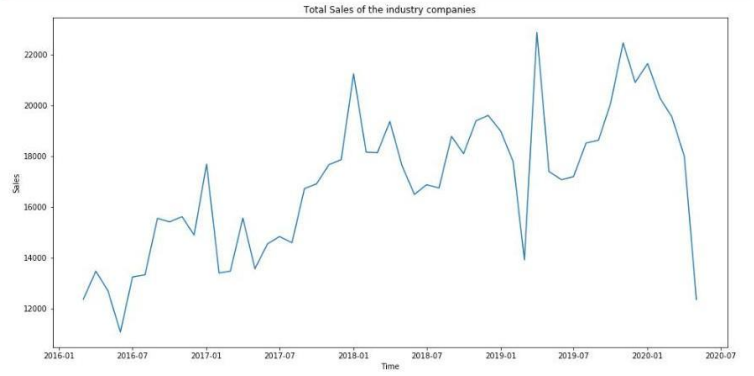

Рисунок 1. Общий объём продаж промышленных предприятий Брестской области

Далее провели сезонное и обычное дифференцирование ряда для того, чтобы добиться стационарности, поскольку критерию Дики-Фуллера не всегда можно доверять.

Для подбора параметров модели построили графики автокорреляционной и частично автокорреляционной функции ряда, как показано на рисунке 3.

```
plt.figure(figsize(10,7))
sm.tsa.seasonal_decompose(ts.item_sales_usd,freq=12, model="additive").plot()
#Проверка на стационарность
print ("Критерий Дики-Фулера: p=%f" % sm.tsa.stattools.adfuller(ts.item_sales_usd)[1])
```

```
Критерий Дики-Фулера: p=0.019627

<Figure size 720x504 with 0 Axes>
```

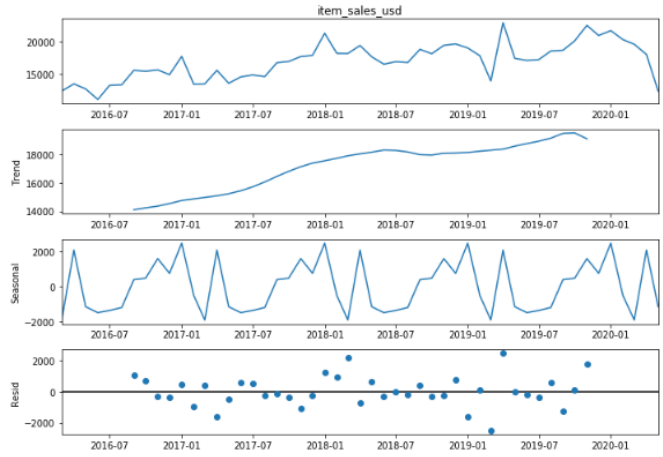

Рисунок 2. STL декомпозиция временного ряда

По первому графику были подобраны начальные приближения для параметров q и Q, по второму графику – приближения для параметров p и P. Дальше сделали перебор всех параметров до приближений, подобранных по графикам автокорреляционной и частично автокорреляционной функции, посчитали для каждого из них значение информационного критерия Акаике, выбрали ту модель, у которой значение этого критерия минимальное.

В результате визуального анализа остатков модели, определили, что в остатках нет никакого тренда, сезонности, они похожи на шум.

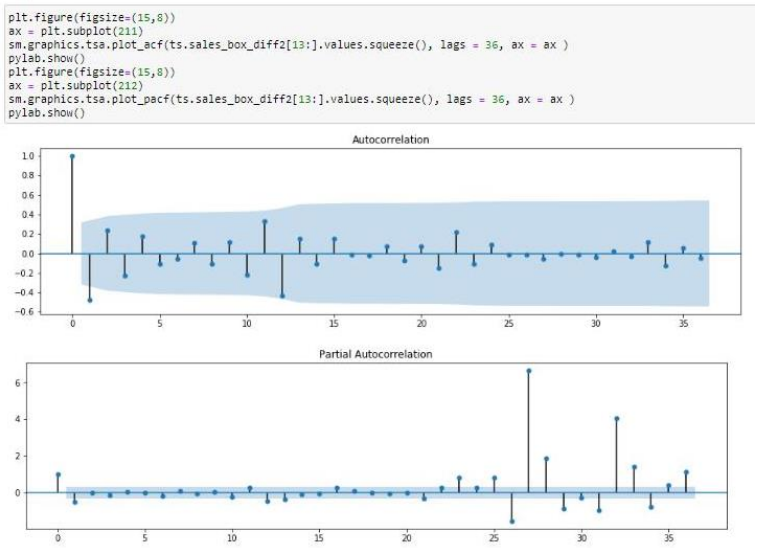

Рисунок 3. Графики корреляционной и частично корреляционной функции ряда

Посмотрели, насколько хорошо эта модель описывает данные, как показано на рисунке 4. Синяя линия – это исходный ряд, красная линия – это то, что предсказывает построенная модель. По графику видно, что модель и данные достаточно похожи.

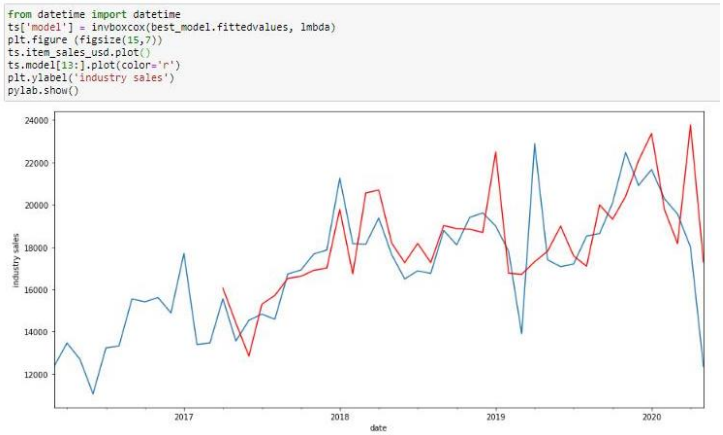

Рисунок 4. Проверка модели

Дальше был построен прогноз на 12 месяцев вперед, он строится при помощи функции predict (Рисунок 5). Синяя линия – это та информация, которая была. Красная линия – это прогноз на 1 год вперед. Он выглядит достаточно адекватным, после плавного подъема следует резкий спад. Этот прогноз передаёт сезонность, которая была выявлена.

Для оценки качества модели отрезали 12 последних наблюдений во временном ряде, построили прогноз и посчитали MAE и MAPE для модели, получили значения MAPE и MAE 11,4 % и 1953,6 тыс. долл. США соответственно. Реальные и прогнозные значения за последние 6 месяцев представлены в таблице 1.

Таблица 1. Реальные и прогнозные значения модели ARIMA

	01-12-2019	01-01-2020	01-02-2020	01-03-2020	01-04-2020	01-05-2020
Реальные значения (тыс. долл. США)	20909,0	21657,1	20276,4	19555,2	17991,0	12349,3
Прогнозные значения (тыс. долл. США)	22077,2	23361,4	19760,1	18164,6	23770,1	17296,5

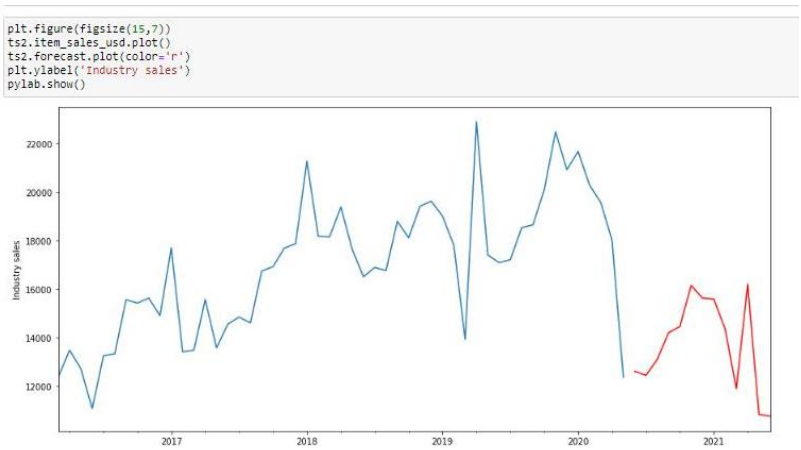

Рисунок 5. Прогноз общего объёма продаж крупных и средних предприятий мебельной отрасли Брестской области

Как видно высокие значения ошибок получили из-за непредвиденного для экономики фактора, реальные и прогнозные значения сильно отличаются по состоянию на 1 апреля и 1 мая 2020 года. Пересчитаем ошибки без последних двух наблюдений, получили значения MAPE и MAE 6,5 % и 1271,7 тыс. долл. США соответственно, что значительно лучше.

Для построения прогноза при помощи библиотеки Fbprophet не нужно настраивать модель, преобразовывать временной ряд, приводить его к стационарности. Для начала нужно привести данные к нужному виду для построения прогноза и подключить библиотеку. Дальше при помощи функции predict строим прогноз на 1 год вперед.

Посмотрели, насколько хорошо эта модель описывает данные (Рисунок 6). По графику видно, что модель и данные достаточно похожи.

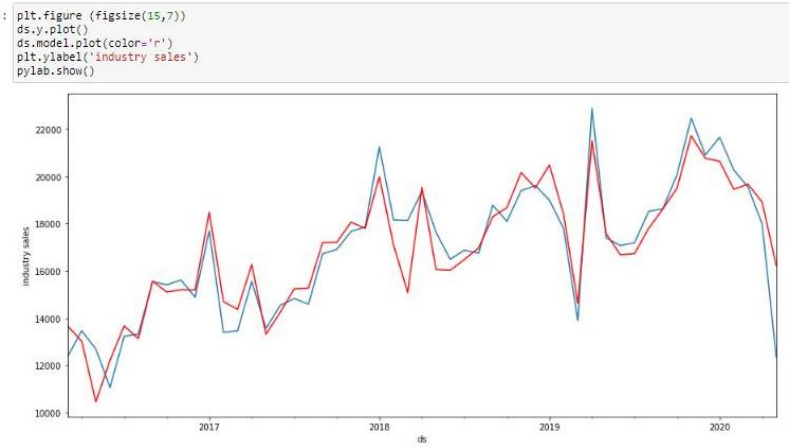

Рисунок 6. Проверка модели

Далее отобразили прогнозные значения на графике на 1 год вперёд (Рисунок 7).

Сравним несколько прогнозных значений с имеющимися реальными, полученными в органах статистики по запросу (Таблица 2).

Таблица 2. Реальные и прогнозные значения модели, построенной при помощи Fbprophet

	01-06-2020	01-07-2020	01-08-2020	01-09-2020	01-10-2020
Реальные значения (тыс. долл. США)	14502,9	19265,8	22965,7	22651,8	25031,3
Прогнозные значения (тыс. долл. США)	17878,9	19225,4	18571,7	20896,4	20328,3

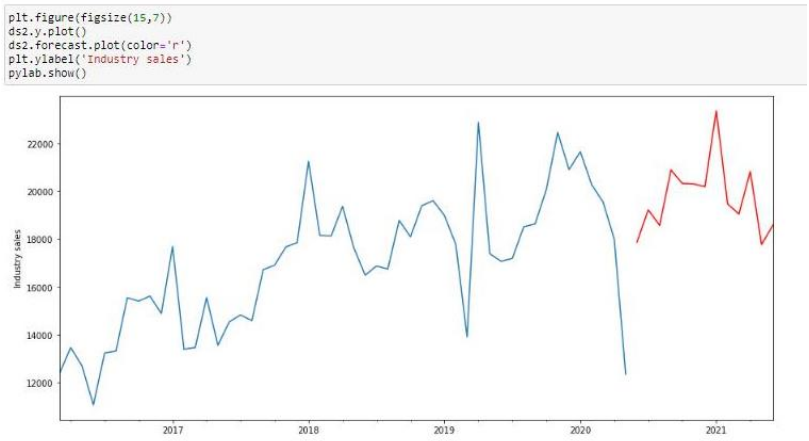

Рисунок 7. Прогноз общего объёма продаж крупных и средних предприятий мебельной отрасли Брестской области

Для оценки качества аналогично отрезали 12 последних наблюдений и построили прогноз, получили значения MAPE и MAE 8,5 % и 1382,3 тыс. долл. США соответственно. Видно, что средняя абсолютная ошибка в процентном выражении и ошибка модели в абсолютных величинах ниже, чем в модели ARIMA. Поэтому можно сделать вывод о том, что конкретно для решения этой задачи больше подходит модель, построенная вторым способом, она является более точной. По таблице 3 видно, что модель хорошо спрогнозировала значения, однако на 1 апреля и 1 мая 2020 года не учла непредвиденный фактор, поэтому ошибки имеют высокие значения. После пересчета ошибок без последних двух наблюдений, получили значения MAPE и MAE 4,6 % и 937,3 тыс. долл. США соответственно, что значительно лучше.

Таблица 3. Реальные и прогнозные значения модели, построенной при помощи Fbprophet

	01-12-2019	01-01-2020	01-02-2020	01-03-2020	01-04-2020	01-05-2020
Реальные значения (тыс. долл. США)	20909,0	21657,1	20276,4	19555,2	17991,0	12349,3
Прогнозные значения (тыс. долл. США)	21421,5	19559,9	19405,1	19273,1	19085,4	18469,6

Для решения задачи прогнозирования объёма продаж предприятий мебельной отрасли Республики Беларусь по отдельным видам продукции был использован язык программирования R. Для решения этой задачи были использованы адаптивные методы прогнозирования. Адаптивные методы прогнозирования позволяют предсказать в краткосрочной перспективе «поведения» показателей. Изначально требуется провести аналитическое выравнивание временного ряда для получения параметров первичной модели, дальше проводится корректировка параметров для выполнения прогнозов [14].

К элементарным адаптивным моделям прогнозирования относятся.

1. Метод Брауна (Brown), простое экспоненциальное сглаживание.
2. Метод Хольта (Holt).
3. Метод Хольта-Винтерса (Holt-Winters).
4. Метод Тейла-Вейджа (Theil-Wage).

Построим прогнозы объёмов продаж при помощи адаптивных методов для одной комбинации предприятия и вида продукции (31.01.12 – Мебель для общественных помещений деревянная) из всех имеющихся данных. Этот временной ряд имеет следующий вид (рисунок 8). Следует отметить, что можно выбрать и другую комбинацию обезличенного предприятия и вида продукции.

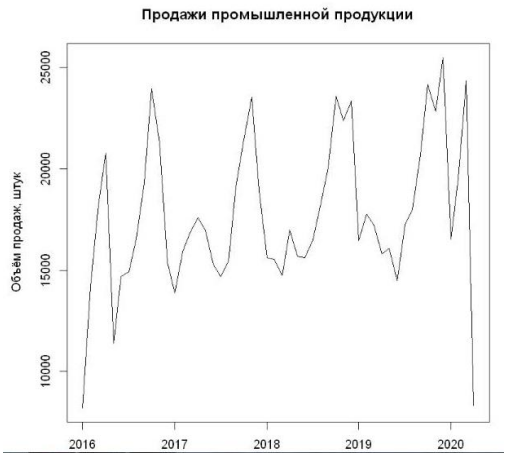

Рисунок 8. Продажи промышленной продукции

Посмотрим на компоненты временного ряда, сделаем STL декомпозицию (рисунок 9). Видим, что временной ряд содержит увеличивающийся тренд, сезонность имеет сложную структуру.

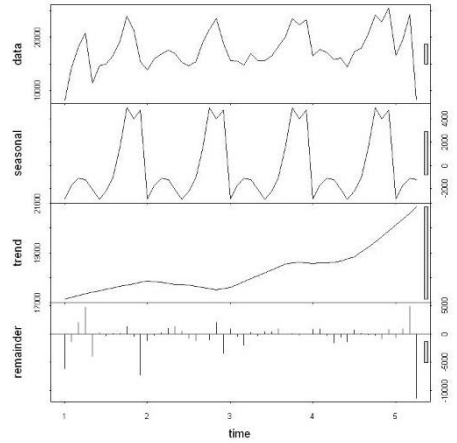

Рисунок 9. STL декомпозиция временного ряда

Очень часто определить вид тренда и сезонности визуально сложно. Поэтому построим несколько адаптивных моделей и выберем из них наилучшую по значениям MAE, MAPE.

Модель Брауна на основе простого экспоненциального сглаживания с автоматически подобранным параметром адаптации, построенная при помощи функции HoltWinters (), представлена на рисунке 10. Здесь отображен прогноз на 12 месяцев вперед. Построенная модель не учитывает ни тренд, ни сезонность, выявленную ранее.

Вторая модель – модель Хольта с линейным трендом и без сезонности с автоматически подобранными параметрами адаптации выглядит так, как показано на рисунке 11. Как видно, эта модель не учитывает выявленную сезонность.

Модель Хольта-Винтерса с линейным трендом и с годовой сезонностью представлена на рисунке 12.

По графику 12 видно, что прогноз для нашего временного ряда выглядит адекватным. Однако для того, чтобы убедиться в адекватности построенных моделей и выбрать из них наилучшую, отрежем 12 последних наблюдений во временном ряде, построим прогноз и посчитаем MAE и MAPE для каждой построенной модели.

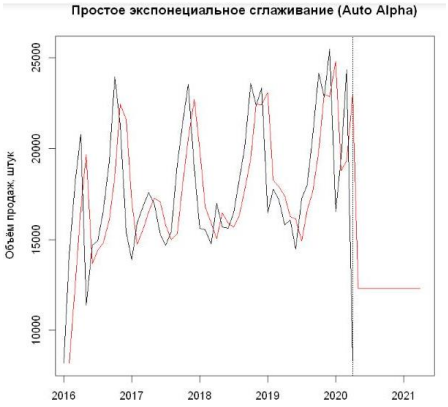

Рисунок 10. График исходного, сглаженных и прогнозных уровней временного ряда модели Брауна

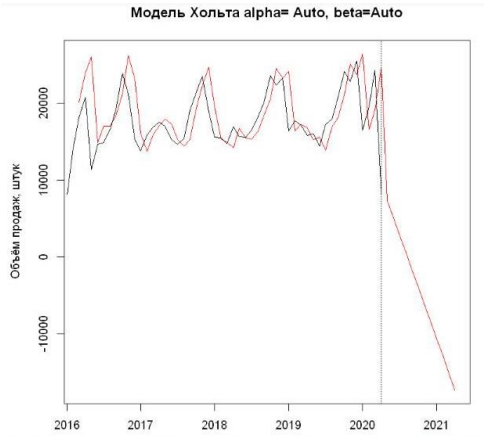

Рисунок 11. График исходного, сглаженных и прогнозных уровней временного ряда модели Хольта

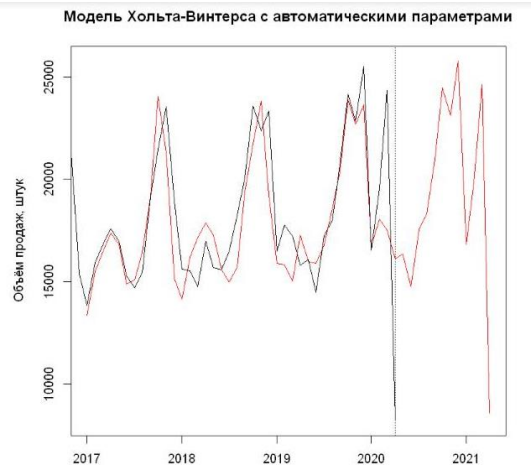

Рисунок 12. График исходного, сглаженных и прогнозных уровней временного ряда модели Хольта-Винтерса

Значения MAE и MAPE для каждой модели представлены в таблице 4.

Таблица 4. Значения MAPE и MAE для построенных адаптивных моделей

Модель Тип ошибки	Модель Брауна (функция HoltWinters())	Модель Хольта (функция HoltWinters())	Модель Хольта-Винтерса (функция HoltWinters())
MAPE (%)	25,6	37,3	14,6
MAE (штук)	4639,4	7899,7	1915,6

Как видно из таблицы, наименьшие значения ошибок получились для модели Хольта-Винтерса, построенной при помощи функции Holt-Winters, с аддитивным трендом и годовой сезонностью, выявленной нами при анализе.

Таким образом, для временного ряда об объёмах продаж комбинации предприятия и кода продукции (31.01.12 – Мебель для общественных помещений деревянная) модель Хольта-Винтерса с аддитивным трендом и годовой сезонностью является наилучшей.

Заключение.

В процессе обработки больших статистических данных технологии DM и ML позволят в целом повысить качество официальной статистики, а также позволят получать новые потенциально полезные знания, выявлять скрытые закономерности и взаимосвязи, получать прогнозирующие результаты, которые будут полезны как государственным структурам, так и бизнесу.

Несмотря на то, что Республика Беларусь активно движется на пути к развитию сферы информатизации и системы электронного правительства, методы интеллектуального анализа данных и машинного обучения еще активно не используются государственными органами, в частности Национальным статистическим комитетом при производстве статистики.

Таким образом, предлагается внедрить подсистему интеллектуального анализа данных в рамках Единой информационной системы государственной статистики и использовать методы DM и ML в при производстве статистики. Поскольку, в отличие от аналитических платформ, где требуются затраты на покупку лицензий, языки Python и R являются бесплатными инструментами для углубленного анализа данных, это внедрение не влечет за собой больших затрат. Помимо этого богатый набор функций стандартной библиотеки Python, широкий набор внешних библиотек, удобный репозиторий пакетов языка R позволяют решать практически любые задачи. Внедрение методов Data Mining в органах статистики позволит проводить глубокий анализ статистических данных, получать новые полезные знания, строить качественные прогнозы, которые можно предоставлять, как на платной, так и на безвозмездной основе, как коммерческим, так и некоммерческим организациям для повышения эффективности их функционирования.

В ходе проведения исследования, для примера были получены данные статистики промышленности в органах государственной статистики и к этим данным применены методы DM и ML. На языке Python были построены прогнозы общего объёма продаж крупных и средних предприятий мебельной отрасли Брестской области методом ARIMA и при помощи библиотеки fbprophet на 1 год вперёд, далее было оценено качество построенных моделей. На языке программирования R были построены прогнозы объёмов продаж одного обезличенного предприятия мебельной отрасли Республики Беларусь по отдельному виду продукции при помощи адаптивных методов, также было оценено качество построенных моделей и выбрана наилучшая.

Список литературы

[1] Степанов, Р.Г. Технология Data Mining: учеб. пособие / Р.Г. Степанов; Казан. Гос. Ун-т им. Ульянова-Ленина. – Казань, 2008. – 57 с.

[2] Технологии Machine Learning, Data Mining [Электронный ресурс] // Российский фонд фундаментальных исследований, 2020. – Режим доступа: http://www.machinelearning.ru/wiki. – Дата доступа: 15.04.2020.

[3] Видеолекции курса «Машинное обучение» [Электронный ресурс] // Школа анализа данных Яндекс, 2020. – Режим доступа : https://yandexdataschool.ru/edu-process/courses/machine-learning. – Дата доступа : 04.01.2020.

[4] Видеолекции курса «Тренды и классификации» [Электронный ресурс] // Новосибирский государственный университет, 2020. – Режим доступа: https://www.coursera.org/learn/trendy-klassifikatsii/home/welcome – Дата доступа: 15.07.2020.

[5] Experimental statistics [Electronic resource] / European statistical system, 2020 – Mode of access : https://ec.europa.eu/eurostat/web/ess/experimental-statistics – Date of access : 20.05.2020.

[6] Statistics Netherlands' innovation portal [Electronic resource] // Statistics Netherlands, 2020 – Mode of

access: https://www.cbs.nl/nl-nl/over-ons/innovatie – Date of access : 23.10.2020.

[7] Customised services & microdata [Electronic resource] // Statistics Netherlands, 2020 – Mode of access: https://www.cbs.nl/en-gb/our-services/customised-services-microdata – Date of access: 18.05.2020.

[8] Experimental statistics [Electronic resource] // Federal Statistical Office Switzerland, 2020 – Mode of access: https://www.experimental.bfs.admin.ch/ – Date of access: 19.05.2020.

[9] Data science for public good https [Electronic resource] // Office for National Statistics, 2021 – Mode of access: datasciencecampus.ons.gov.uk/ – Date of access: 19.01.2021.

[10] Официальная статистика [Электронный ресурс] // Национальный статистический комитет, 2020. – Режим доступа: https://www.belstat.gov.by/. – Дата доступа: 22.04.2020.

[11] Закон Республики Беларусь «О государственной статистике» [Электронный ресурс] // Национальный статистический комитет Респ. Беларусь. – Минск, 2004. – Режим доступа: https://www.belstat.gov.by/o-belstate_2/pravovye-osnovy-gosudarstvennoi-statistiki-respubl/zakon-respubliki-belarus-o-gosudarstvennoi-statist/ – Дата доступа: 21.04.2020.

[12] Видеолекции курса «Прикладные задачи анализа данных» [Электронный ресурс] // Московский физико-технический, Яндекс, 2020. – Режим доступа: https://www.coursera.org/learn/data-analysis-applications?authMode=login – Дата доступа: 01.08.2020.

[13] Предсказываем будущее с помощью библиотеки Facebook Prophet [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/ods/blog/323730/. – Дата доступа: 04.08.2020.

[14] Атчаде М. Н. Адаптивные методы прогнозирования: реализация в Excel и программе R: учебное пособие / Под ред. И. И. Елисеевой. – СПб. : СПбГЭУ, 2018. – 101 с.

USE OF DATA MINING TECHNOLOGIES WITHIN THE UNIFIED STATE STATISTICS INFORMATION SYSTEM

A.A.KOROBCHUK
Main specialist of
General Statistical Office of the Brest region

General Statistical Office of the Brest region
E-mail: alinakorobchuk94@mail.ru

Abstract. Statistical information is one of the basic components of state information resource. Today there are higher requirements for statistical information quality. Use of data mining and machine learning technologies by statistical offices within the unified state statistics information system improves the quality of official statistics and createsinnovative statistics. Use of these technologies by foreign statistical offices was considered. Analysis of industrial statistics was done and predictive models were created by different time series analysis methods in the programming languages Python and R, the quality of created models was assessed.

Keywords: statistical data, data mining, machine learning, experimental statistics, time series analysis methods, forecasting.