УДК 004.043

Выявление зависимости количества проданных легковых автомобилей в России от цен на нефть

Маринчук Александр Сергеевич
Приамурский государственный университет им. Шолом-Алейхема
Студент

Баженов Руслан Иванович
Приамурский государственный университет им. Шолом-Алейхема
к.п.н., доцент, зав. кафедрой информационных систем, математики и правовой информатики

Аннотация
В данной статье рассматривается зависимость между количеством проданных легковых автомобилей в России и ценами на нефть. Был сделан выбор интеллектуального метода, и построена по нему модель, смотря на которую определили в какую сторону направлен автомобильный рынок и смогли спрогнозировать его состояние в будущем.
Ключевые слова: Язык программирования R, интеллектуальный метод, построение модели, прогноз.

Identifying the dependence of the number of cars sold in Russia on oil prices

Marinchuk Alexander Sergeevich
Sholom-Aleichem Priamursky State University
Student

Bazhenov Ruslan Ivanovich
Sholom-Aleichem Priamursky State University
Candidate of pedagogical sciences, associate professor, Head of the Department of Information Systems, Mathematics and Law Informatics

Abstract
This article examines the relationship between the number of cars sold in Russia and the price of oil. A choice was made of the intellectual method, and a model was built on it, looking at which we saw in which direction the automobile market was directed and could predict its condition in the future.
Keywords: Programming language R, intelligent method, model building, forecast

В настоящее время трудно предугадать как сложится ситуация на автомобильном рынке России. Чтобы быть подготовленным к разного рода событиям следует выяснить факторы от которых зависит число проданных легковых автомобилей и спрогнозировать количество купленных россиянами

легковых машин. Уже несколько лет просматривается некая зависимость между количеством проданных легковых автомобилей в России и ценами на нефть. Выбрав интеллектуальный метод и построив по нему модель, мы увидим в какую сторону направлен автомобильный рынок и сможем спрогнозировать его состояние в будущем.

Целью нашего исследования является создание предсказательных моделей для выявления зависимости количества проданных легковых автомобилей в России от цен на нефть и выбор из этих моделей наилучшей для построения правдивого прогноза.

В статье В.Н. Беговщиц и др. использован метод линейной регрессии для обработки данных нестационарного аэродинамического эксперимента[1]. О.В. Прокофьев, И.Ю.Семочкина провели анализ возможностей применения языка программирования R и среды RStudio для математической обработки данных[2]. М.И. Калугина, С.В. Бегичева рассмотрели различные способы визуализации больших массивов данных в среде R с использованием следующих библиотек: ggplot2, RGraphics, grid, gridExtra, GGally[3]. В своей статье И.А. Лызин описал важность визуализации данных с помощью языка программирования R[4]. В статье R. Ihaka, R. Gentleman рассмотрена библиотека Dendrochronology Program в R (dplR), которая представляет собой пакет для обработки и анализа данных [5]. A. Bunn, M. Korpela обсудили опыт разработки и внедрения статистического компьютерного языка R в своей статье[6]. H. M. Wagner рассмотрел и описал метод регрессионного анализа [7].

Данными для нашего исследования будут являться количество проданных легковых автомобилей в России и цены на нефть за период с 2002 по 2016 год. Данные были взяты со свободной энциклопедии «Википедия», которая является источником, подкрепленным множеством документов по нужной тематике.

Для того чтобы выявить зависимость между количеством проданных автомобилей в России и ценами на нефть построим модель по методу линейной регрессии. Вторым и третьим методом для исследования будут являться соответственно полиномиальные регрессии второй и третьей степени. Сравнивая после построения данные модели через определенные критерии, мы выберем ту, которая покажет себя более адекватной в плане прогноза.

Модели, которые мы хотим построить, можно легко реализовать на языке программирования R, который предназначен для статистического анализа и обработки данных. Для работы с языком R будем использовать свободную среду разработки программного обеспечения под названием «RStudio». Благодаря ряду своих особенностей этот активно развивающийся программный продукт делает работу с R очень удобной. R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей

применения. В базовую поставку R включен основной набор пакетов, а всего по состоянию на 2017 год доступно более 11778 пакетов

Прежде чем приступать к работе в «RStudio» следует описать применяемые методы уравнениями для того, чтобы лучше понять работу модели. Математическое уравнение, которое оценивает линию простой линейной регрессии, выглядит следующим образом: $Y = a + bx$.

X называется независимой переменной или предиктором.

Y – зависимая переменная или переменная отклика. Это значение, которое мы ожидаем для Y (в среднем), если мы знаем величину X, т.е. это «предсказанное значение Y»

Полиномиальная регрессия означает приближение данных (X_I, Y_i) полиномом k-й степени

$$Y(x) = a + bx + cx^2 + dx^3 + \ldots + hx^k.$$

При k = 1 полином описывают прямой линией, при k = 2 – параболой, при k = 3 – кубической параболой и т.д.

Теперь приступим к использованию самой программы «RStudio» и начнем с того, что высчитаем коэффициент корреляции, который покажет зависимость между нашими данными. Математически он задан уравнением на рис.1.

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2\sum_{i=1}^{m}(y_i-\bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}},$$

Рис 1 – Уравнение коэффициента корреляции Пирсона

На языке программирования R коэффициент корреляции высчитывается через функцию «cor».

```
> #Подключаем необходимые для работы библиотеки.
> library(DAAG)
> library(corrplot)
> library(ggplot2)
> library(cvTools)
> #Загружаем данные в RStudio из файла.
> zav <- read.table("dann1.txt", header = TRUE )
> #Высчитываем коэффициент корреляции Пирсона(где Neft и Avt названия столбцов наших данных).
>  cor(Avt,Neft)
[1] 0.88539
```

Рис 2 – Результат выполнения функции «cor» для линейной регрессии

Как видно на рисунке 2 показан результат исполнения функции «cor» между нашими данными, где коэффициент корреляции равен 0.88539, это означает, что, если данный коэффициент приближен к 0.8 или выше его, то линейная регрессия этих зависимых данных имеет хорошую связь. Аналогично считаем коэффициенты корреляции для полиномиальных регрессий.

```
> cor(Avt,Neft + I(Neft^2))
[1] 0.9005849
> cor(Avt,Neft + I(Neft^2)+I(Neft^3))
[1] 0.8965814
>
```

Рис 3 – Результат выполнения функции «cor» для полиномиальных регрессий

Смотря на рисунок 3 можно увидеть еще более крепкую связь между данными при методах полиномиальных регрессий. Построим график зависимости количества проданных машин в России от цен на нефть и посмотрим визуально на связь между данными, чтобы в дальнейшем сравнивать с построенными графиками моделей.

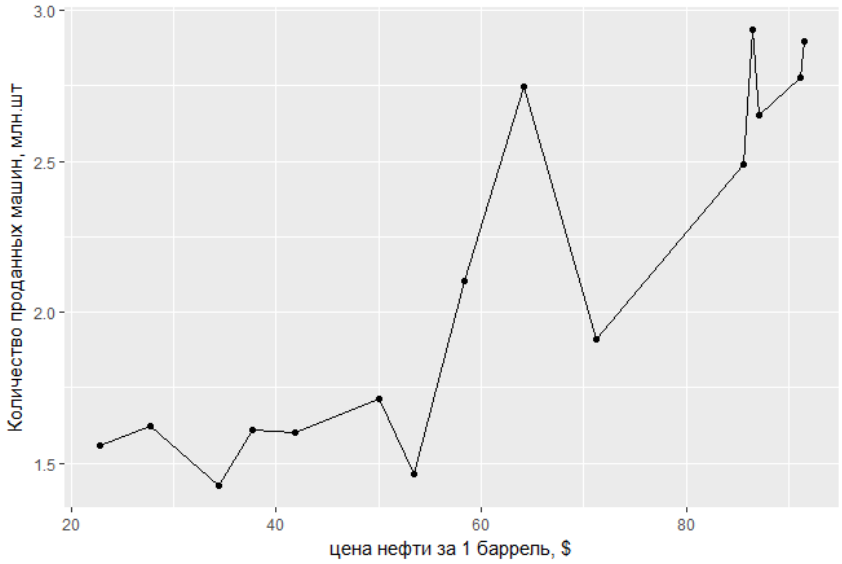

Рис 4 – График зависимости

Смотря на рисунок 4 можно явно увидеть линейную зависимость между ценами на нефть и количеством проданных автомобилей в России, но, чтобы убедиться какая из рассматриваемых нами моделей более адекватна, следует рассмотреть каждую модель с точки зрения определенных критериев.

В нашем исследовании есть лишь один предиктор – цены на нефть за определенный период и как правило, чем больше предикторов мы включаем в модель, тем более точными становятся предсказания. Однако при избыточном усложнении модели качество ее предсказаний начнет снижаться в связи с явлением, известным как переобучение модели. Для борьбы с переобучением используют разнообразные методы отбора информативных предикторов и перекрестную проверку ("cross-validation").

Одним из наиболее распространенных подходов при построении предсказательных моделей является выполненное случайным образом разбиение исходных данных на обучающую и проверочную выборки. Как следует из названия, обучающая выборка используется для "обучения" той или иной модели, т.е. для построения математических отношений между

переменной-откликом и предикторами, тогда как проверочная выборка применяется для оценки предсказательных свойств модели на данных, которые она (модель) раньше "не видела". Более того, при наличии небольшого числа наблюдений разбиение исходных данных на обучающую и проверочную выборки не всегда является возможным. В таких случаях используется перекрестная проверка с последовательным исключением одного наблюдения.

Разобьем наши данные на обучающие и проверочные для построения моделей.

```
> #Разбиваем данные на обучающие и проверочные.
>  # обучающие данные:
> sat1 <- zav[-c(14,15),-1]
> # проверочные данные:
> sat2 <- zav[c(14,15),-1]
```

Рис 5 – Разбиение данных на обучающие и проверочные

Теперь непосредственно построим линейную модель и модели полиномов второй и третьей степени.

```
> set.seed(101) # для воспроизводимости результата.
> M1 = lm(Avt ~ Neft, sat1) #  линейная модель.
> M2 = lm(Avt ~ Neft+I(Neft^2), sat1) # полином 2-й степени.
> M3 = lm(Avt ~ Neft+I(Neft^2)+I(Neft^3), sat1)# полином 3-й степени.
```

Рис 6 – Построение моделей

Для выбора оптимальной модели необходимо определиться с критерием, который позволит выполнить объективное сравнение качества альтернативных моделей. При решении регрессионных задач, когда зависимой является количественная переменная, такого рода критерии, как правило, основаны на остатках, т.е. разницах между наблюдаемыми и предсказанными значениями отклика. В частности, распространенным критерием является квадратный корень из среднеквадратичной ошибки (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Рис 7 – Критерий RMSE

Как видно из приведенной формулы, RMSE показывает, насколько велика, в среднем, разница между действительными наблюдениями и значениями, предсказанными моделью. По определению, чем меньше RMSE, тем точнее предсказания.

Рассчитаем критерий RMSE для наших моделей.

```
> RMSE1 = sqrt(sum((sat1$Avt - predict(M1))^2)/nrow(sat1))#Критерий RMSE для линейной модели.
> RMSE2 = sqrt(sum((sat1$Avt - predict(M2))^2)/nrow(sat1))#Критерий RMSE для полинома 2-й степени.
> RMSE3 = sqrt(sum((sat1$Avt - predict(M3))^2)/nrow(sat1))#Критерий RMSE для полинома 3-й степени.
> RMSE1
[1] 0.2749008
> RMSE2
[1] 0.2561422
> RMSE3
[1] 0.2531876
```

Рис 8 – Критерии RMSE

Из полученных результатов можно заключить, что кубическая модель лучше подходит для предсказания количества проданных автомобилей в России, чем простая линейная регрессии. Здесь, однако, следует отметить очень важный момент: величины RMSE1, RMSE2 и RMSE3 были рассчитаны по обучающим данным, т.е. по тем же данным, которые были использованы для подгонки моделей M1, M2 и M3. Поскольку метод наименьших квадратов, на основе которого происходила подгонка этих моделей, пытается (опосредованно) минимизировать RMSE, приведенные оценки RMSE1, RMSE2 и RMSE3 могут оказаться слишком оптимистичными характеристиками предсказательных свойств моделей M1, M2 и M3 (эта проблема особенно актуальна для сложных моделей, включающих большое число параметров и, как следствие, рискующих быть переобученными).

Для решения указанной проблемы (т.е. для получения несмещенных оценок предсказательных свойств моделей) применяются методы формирования повторных выборок, при помощи которых обучающие данные случайным образом разбиваются на несколько частей. Эти части затем попеременно используются для подгонки разных версий одной и той же модели. Каждый раз данные, не участвующие в подгонке конкретной версии модели, используются для расчета соответствующего критерия качества (например, RMSE). Полученные таким образом значения этого критерия далее усредняются, что позволяет получить несмещенную оценку предсказательных свойств модели. Наиболее распространенным способом получения подобных несмещенных оценок критериев качества является упомянутая выше перекрестная проверка.

Выполним 5-кратную перекрестную проверку для нашего примера. Для этого воспользуемся функцией «cvLm», которая по умолчанию использует RMSE как критерий качества модели.

```
> cvLm(M1, K = 5)
5-fold CV results:
       CV
0.3399307
> cvLm(M2, K = 5)
5-fold CV results:
      CV
0.315366
> cvLm(M3, K = 5)
5-fold CV results:
       CV
0.3417346
```

Рис 9 – Перекрестная проверка

Полученные результаты говорят нам о том, что предпочтение следует отдать модели M2. Однако обратите внимание на то, что значения RMSE для моделей оказались выше тех, которые мы рассчитали ранее. Этот пример наглядно иллюстрирует, что оценка качества модели, полученная на тех же данных, которые использовались для подгонки самой этой модели, практически всегда является чрезмерно оптимистичной.

Окончательный выбор модели среди нескольких альтернатив, отобранных на предыдущем шаге, выполняют путем расчета критерия качества по данным из проверочной выборки. Для сравнения выполним расчет по всем моделям.

```
> sqrt(sum((sat2$Avt - predict(M1, sat2))^2)/nrow(sat2))
[1] 0.1593528
> sqrt(sum((sat2$Avt - predict(M2, sat2))^2)/nrow(sat2))
[1] 0.1472961
> sqrt(sum((sat2$Avt - predict(M3, sat2))^2)/nrow(sat2))
[1] 0.09394682
```

Рис 10 – Критерий RMSE для проверочных данных

Таким образом, выбор следует сделать в сторону полиномиальной модели второй степени, так как эта модель с большей вероятностью будет давать более точные данные, чем линейная модель и полиномиальная модель третьей степени

Графически представим наши модели для наглядности, используя функцию «ggplot» и сравнивая полученные графики с рисунком 4.

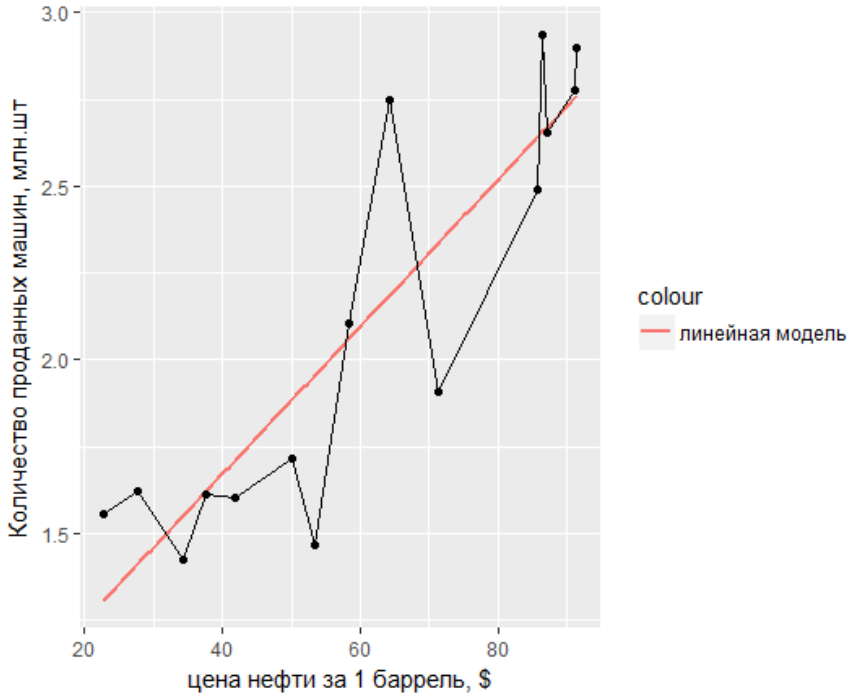

Рис 11 – Линейная модель

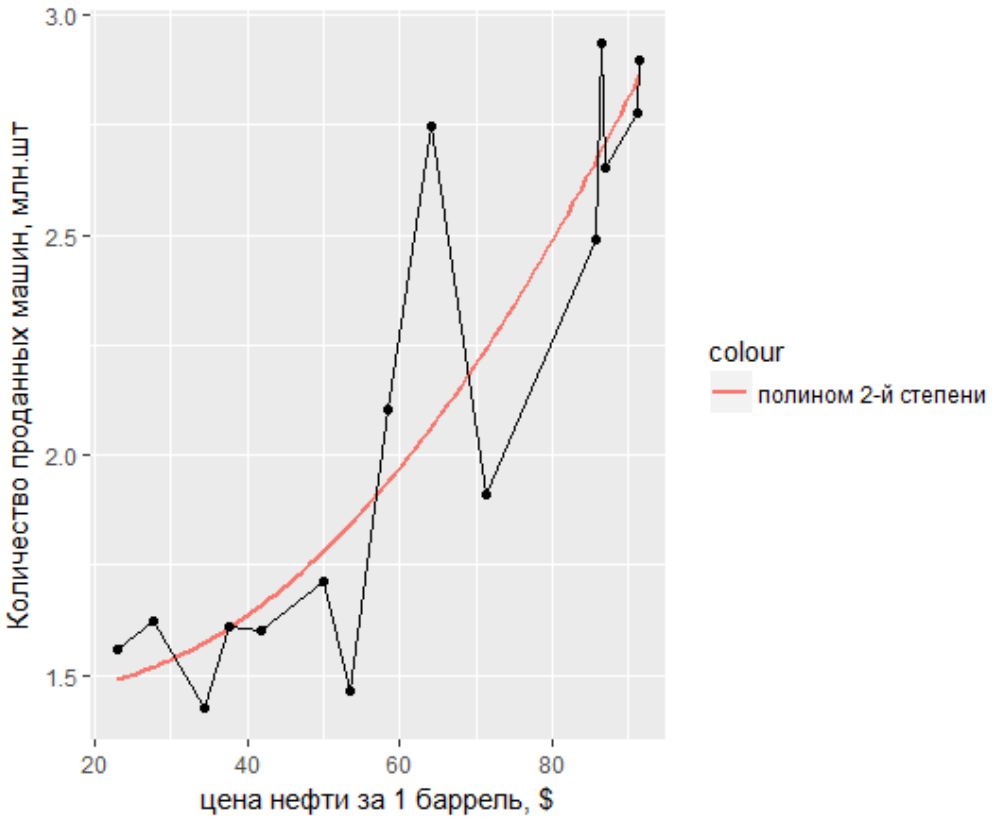

Рис 12 – Полиномиальная модель 2-й степени

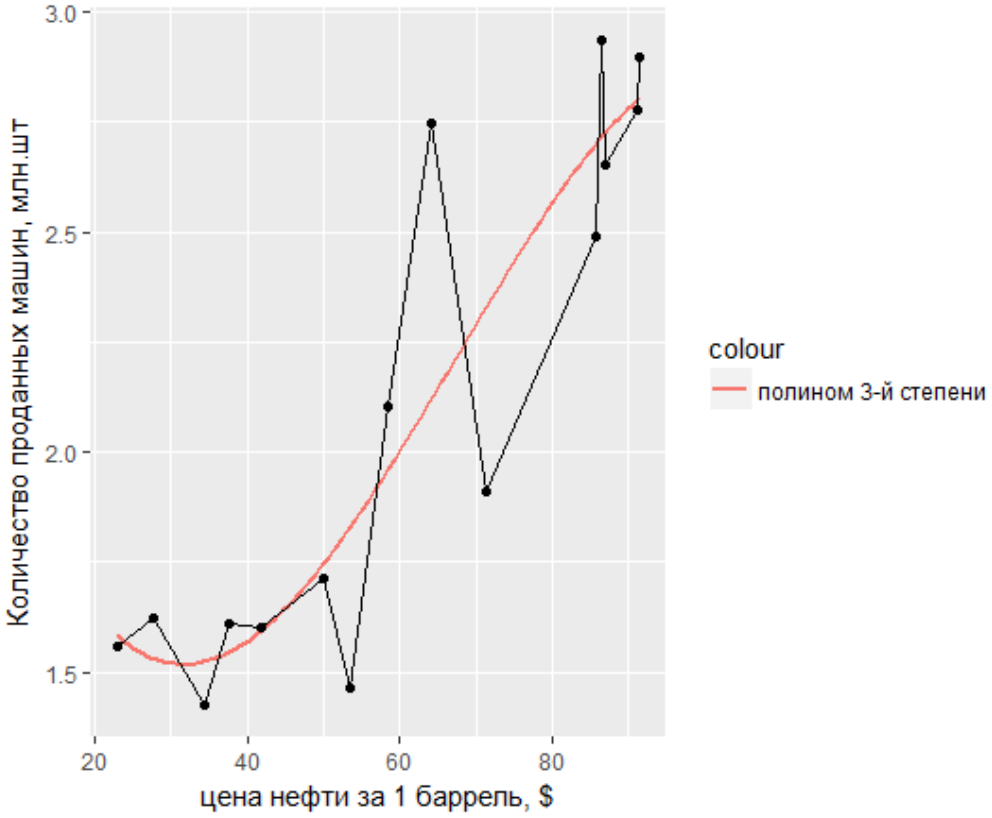

Рис 13 – Полиномиальная модель 3-й степени

Таким образом, в статье была достигнута цель с построением наиболее адекватной модели для выявления зависимости количества проданных

автомобилей в России от цен на нефть в системе R. Если судить по графикам, то хочется отдать предпочтение полиномиальной модели третьей степени, но после вычисления критерия RMSE для каждой модели и перекрестной проверки становится понятным, что для прогноза следует отдать предпочтение полиномиальной модели второй степени. График наиболее адекватной модели дает нам будущее представление о том, что цены на нефть будут расти, как и количество продаваемых автомобилей в России, но возможны скачки в определенные периоды времени.

Библиографический список

1. Беговщиц В.Н., Колинько К.А., Миатов О. Л, Храбров А.Н. Использование метода линейной регрессии для обработки данных нестационарного аэродинамического эксперимента // Ученые записки Цаги. 1996. №3-4 . С. 30-38.
2. Прокофьев О. В., Семочкина И. Ю. Применение языка R и среды RStudio для математической обработки данных // Современные информационные технологии. 2017 . №25. С. 47-51.
3. Калугина М.И., Бегичева С.В. Современные возможности визуализации результатов исследований в среде R // BI-технологии и корпоративные информационные системы в оптимизации бизнес-процессов. Екатеринбург: Уральский государственный экономический университет, 2016. С. 51-55.
4. Лызин И.А. Использование визуализации при анализе медицинских данных // Информационные технологии в науке, управлении, социальной сфере и медицине. Томск: Национальный исследовательский Томский политехнический университет, 2017. С. 339-342.
5. Ihaka R., Gentleman R. R: a language for data analysis and graphics //Journal of computational and graphical statistics. 1996. Т. 5. №. 3. С. 299-314.
6. Bunn A., Korpela M. R: A language and environment for statistical computing. 2013.
7. Wagner H. M. Linear programming techniques for regression analysis //Journal of the American Statistical Association. 1959. Т. 54. №. 285. С. 206-212.
8. Рейтинги продаж автомобилей в России в 2017 - 2018 году // http://serega.icnet.ru/cars-sales-actual-russia.html (дата обращения: 20.04.2018).
9. Использование языка R для статистической обработки данных // http://kpfu.ru/docs/F407025247/metodichka_R_2.pdf (дата обращения: 20.04.2018).
10. Классические методы статистики: коэффициент корреляции // http://r-analytics.blogspot.ru/2012/09/blog-post_6280.html#.WtttX8iFPIW (дата обращения: 20.04.2018).
11. Создание предсказательных моделей: основные шаги // http://r-analytics.blogspot.ru/2015/05/blog-post.html#.WtttfsiFPIV (дата обращения: 20.04.2018).

12.Цены на нефть // https://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BD%D1%8B_%D0%BD%D0%B0_%D0%BD%D0%B5%D1%84%D1%82%D1%8C (дата обращения: 20.04.2018).