Моделирование и анализ данных
2020. Том 10. № 2. С. 39–48
DOI: https://doi.org/10.17759/mda.2020100203
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)


Modelling and Data Analysis
2020. Vol 10. no. 2. pp. 39–48
DOI: https://doi.org/10.17759/mda.2020100203
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)


## АНАЛИЗ ДАННЫХ

УДК 519.246.8; 37.018.432

# Прогнозирование необходимого онлайн-школе числа преподавателей

***Сологуб Г.Б.****
Московский авиационный институт
(национальный исследовательский университет)
(ФГБОУ ВО МАИ (НИУ)), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0002–5657–4826
e-mail: glebsologub@ya.ru

***Пухов В.А.*****
Московский авиационный институт
(национальный исследовательский университет)
(ФГБОУ ВО МАИ (НИУ)), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0002–8078–6386
e-mail: csguard26@gmail.com

***Цыпленков Л.С.*******
Московский авиационный институт
(национальный исследовательский университет)
(ФГБОУ ВО МАИ (НИУ)), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0001–6010–1223
e-mail: leonidtsyplenkov@gmail.com

**Для цитаты:**
*Сологуб Г.Б., Пухов В.А., Цыпленков Л.С.* Прогнозирование необходимого онлайн-школе числа преподавателей // Моделирование и анализ данных. 2020. Том 10. № 2. С. 39–48. DOI:10.17759/mda.2020100203

***Сологуб Глеб Борисович,** кандидат физико-математических наук, доцент кафедры математической кибернетики института «Информационные технологии и прикладная математика» Московского авиационного института (национального исследовательского университета), г. Москва, Россия, ORCID: https://orcid.org/0000–0002–5657–4826, e-mail: glebsologub@ya.ru

****Пухов Вячеслав Александрович,** студент бакалавриата института «Информационные технологии и прикладная математика» Московского авиационного института (национального исследовательского университета), г. Москва, Россия, ORCID: https://orcid.org/0000–0002–8078–6386, e-mail: csguard26@gmail.com

*****Цыпленков Леонид Сергеевич,** студент бакалавриата института «Информационные технологии и прикладная математика» Московского авиационного института (национального исследовательского университета), г. Москва, Россия, ORCID: https://orcid.org/0000–0001–6010–1223, e-mail: leonidtsyplenkov@gmail.com

В статье описан подход к построению прогноза числа преподавателей, которое необходимо нанять онлайн-школе английского языка, на основе анализа исторических данных о проведенных уроках с использованием моделей линейной регрессии и тройного экспоненциального сглаживания.

***Ключевые слова:*** временные ряды, анализ временных рядов, прогнозирование временных рядов, онлайн-образование.

## 1. ВВЕДЕНИЕ

Онлайн-образование становится всё более и более популярным, поэтому число учеников в онлайн-школах растёт. В этой работе мы рассматриваем онлайн-школу английского языка, в которой за несколько лет с момента основания число учащихся достигло более 100 тысяч и продолжает год к году кратно расти. Каждый день новые ученики регистрируются на сайте школы, совершают оплату и начинают заниматься. Вместе с тем существует и естественный отток учеников – со временем ученики перестают заниматься. В результате количество учеников всё время меняется. Аналогично, меняется и число учителей – новые учителя начинают работу, а некоторые старые уходят. Для обеспечения возрастающего спроса школе необходимо привлекать всё большее число учителей. Для правильного планирования бизнес-процесса привлечения учителей требуется прогноз необходимого числа учителей по меньшей мере на следующий месяц.

## 2. ПОСТАНОВКА ЗАДАЧИ

Исходные данные представляют собой набор записей обо всех уроках, проведенных в онлайн-школе за 25 месяцев, начиная с сентября 2017-го года и заканчивая сентябрем 2019-го года, в виде таблицы в формате csv, состоящей из трех столбцов: дата и время начала занятия, id учителя, id ученика.

Необходимо построить прогноз числа учителей, требуемых онлайн-школе в октябре 2019-го года.

## 3. ПРОГНОЗИРОВАНИЕ КОЛИЧЕСТВА ПРЕПОДАВАТЕЛЕЙ НА ОСНОВЕ МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ СТАТИСТИК CHURN RATE И INCOME RATE

Сформулируем алгоритм построения модели.

1. Выполнить предобработку данных путем вычисления количества учителей в каждом месяце.
2. Построить линейную регрессию по числу учителей (тренд).
3. Вычислить статистики churn rate и income rate на каждый месяц.
4. Построить прогноз с использованием значений тренда и этих статистик.

Сначала из исходных данных вычислим количество работавших учителей по месяцам, считая, что преподаватель работал в определенном месяце, если он провел в нем хотя бы один урок. Получим таблицу 1, содержащую 25 величин.

Таблица 1

**Количество учителей по месяцам**

| Год | Месяц | Количество учителей |
|---|---|---|
| 2017 | Сентябрь | 1207 |
| | Октябрь | 1273 |
| | Ноябрь | 1463 |
| | Декабрь | 1693 |
| 2018 | Январь | 1806 |
| ... | ... | ... |
| 2019 | Май | 7024 |
| | Июнь | 7154 |
| | Июль | 7303 |
| | Август | 7776 |
| | Сентябрь | 8252 |

Затем построим линейную регрессию. Таким образом, мы получим “тренд” – общую тенденцию роста числа учителей.

Уравнение линейной регрессии для $n$ наблюдений имеет вид:

$$Y_i = a_0 + kt_i, i = 1, 2, \ldots, n,$$

где $t_i \in \{0\} \cup \mathbb{N}$ независимая переменная, $Y_i \in \mathbb{Z}$ - зависимая переменная, которую будем предсказывать (число преподавателей в конце *i*-го месяца), $a_0 \in \mathbb{R}$ – значение $Y_i$ при $t_i = 0, k \in \mathbb{R}$ – величина, на которую в среднем увеличивается $Y_i$, если увеличить *t* на единицу.

Запишем уравнение линейной регрессии в матричной форме:

$$Y = A \cdot \theta,$$

где $Y$ – вектор наблюдений; $A$ – матрица, в которой первый столбец единичный, а второй – значения параметра $t, \theta = (a_0, k)^T$.

Вектор исследуемых параметров оценивается методом наименьших квадратов:

$$\hat{\theta} : (Y - A \cdot \theta)^2 \to \min_{\theta \in R^2},$$

$$\hat{\theta} = \arg\min_{\theta \in R^2} (Y - A \cdot \theta)^2 = (A^T A)^{-1} A^T Y.$$

В рассматриваемом случае наблюдения – это количество преподавателей, независимая переменная – номер месяца по счету.

На рисунке 1 построена прямая – график линейной регрессии – и изображены точки – фактическое число преподавателей в каждый месяц.

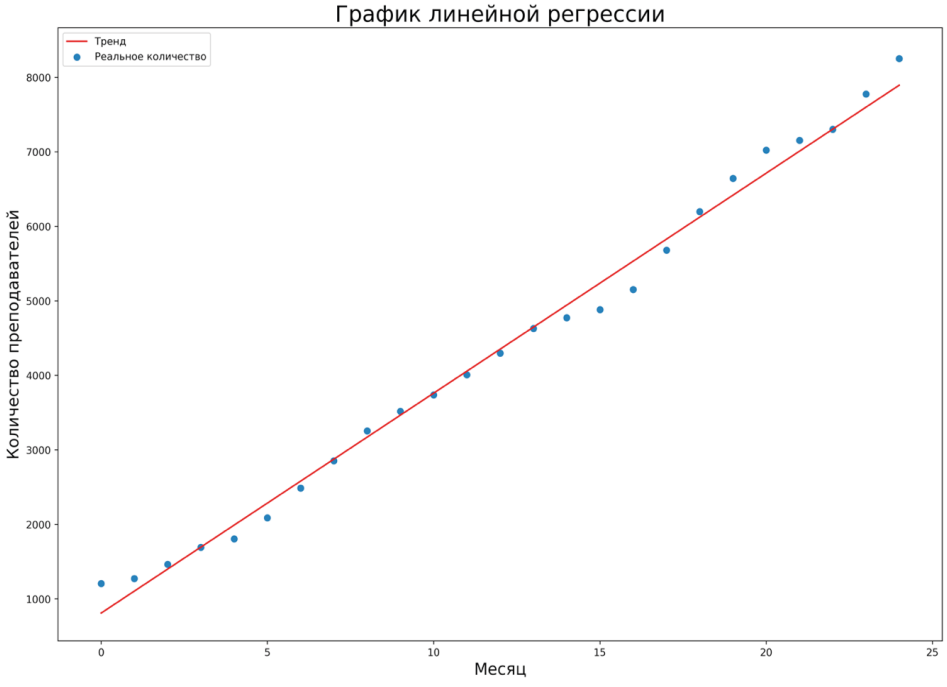


*Рис.1. График линейной регрессии. На оси абсцисс – номер месяца, на оси ординат – количество преподавателей в этом месяце. Красная линия – график линейной регрессии; синие точки – исторические данные*

Далее вводится сезонность изменений с помощью **churn rate** и **income rate**, или коэффициентов оттока и притока.

Строится Таблица 2, в которой для каждого доступного месяца считается:

1) количество преподавателей,
2) сколько из них осталось с предыдущего месяца,
3) сколько из них ушло,
4) сколько пришло новых учителей.

Замечания.

- Количество преподавателей – это число преподавателей, которые провели хотя бы одно занятие в этом месяце.
- Если преподаватель провел хотя бы один урок в предыдущем месяце и провел хотя бы один урок в этом, считается, что он остался работать.
- Если же в предыдущем месяце преподаватель давал занятия, а в этом – нет, считается, что он ушел.
- Если преподаватель впервые провел урок на платформе, либо же у него был перерыв от деятельности в течение более чем одного месяца, он считается новым.

Таблица 2

**Количество учителей, вычисленное для нахождения коэффициентов притока и оттока**

| Год | Месяц | Количество учителей в месяце | Количество оставшихся с прошлого месяца учителей | Количество ушедших с прошлого месяца учителей | Количество новых учителей в этом месяце |
|---|---|---|---|---|---|
| 2017 | Сентябрь | 1207 | Неизвестно | Неизвестно | Неизвестно |
| | Октябрь | 1273 | 1140 | 67 | 133 |
| | Ноябрь | 1463 | 1202 | 71 | 261 |
| | Декабрь | 1693 | 1381 | 82 | 312 |
| 2018 | Январь | 1806 | 1570 | 123 | 236 |
| … | … | … | … | … | … |
| 2019 | Май | 7024 | 6299 | 345 | 725 |
| | Июнь | 7154 | 6553 | 471 | 601 |
| | Июль | 7303 | 6544 | 610 | 759 |
| | Август | 7776 | 6756 | 547 | 1020 |
| | Сентябрь | 8252 | 7166 | 610 | 1086 |

Пусть

- $Count(n)$ – количество учителей в $n$-м месяце;
- $Left(n)$ – количество ушедших учителей в $n$-м месяце;
- $New(n)$ – количество новых учителей в $n$-м месяце;
- $Churn(n)$ – churn rate в $n$-м месяце;
- $Income(n)$ – income rate в $n$-м месяце.

Тогда:

$$Churn(n) = \frac{Left(n)}{Count(n-1)},$$

$$Income(n) = \frac{New(n)}{Count(n-1)}.$$

Рассчитаем коэффициенты притока и оттока (см. Таблицу 3).

Таблица 3

**Расчет коэффициентов притока и оттока**

| Год | Месяц | Кол-во учителей | Осталось учителей | Ушло учителей | Новых учителей | К-т оттока | К-т притока |
|---|---|---|---|---|---|---|---|
| 2017 | Сентябрь | 1207 | | | | | |
| | Октябрь | 1273 | 1140 | 67 | 133 | 0,0555 | 0,1101 |
| | Ноябрь | 1463 | 1202 | 71 | 261 | 0,0557 | 0,2050 |

| Год | Месяц | Кол-во учителей | Осталось учителей | Ушло учителей | Новых учителей | К-т оттока | К-т притока |
|---|---|---|---|---|---|---|---|
| | Декабрь | 1693 | 1381 | 82 | 312 | 0,0560 | 0,2132 |
| 2018 | Январь | 1806 | 1570 | 123 | 236 | 0,0726 | 0,1393 |
| … | … | … | … | … | … | … | … |
| 2019 | Май | 7024 | 6299 | 345 | 725 | 0,0519 | 0,1091 |
| | Июнь | 7154 | 6553 | 471 | 601 | 0,0670 | 0,0855 |
| | Июль | 7303 | 6544 | 610 | 759 | 0,085 | 0,1060 |
| | Август | 7776 | 6756 | 547 | 1020 | 0,0749 | 0,1396 |
| | Сентябрь | 8252 | 7166 | 610 | 1086 | 0,0784 | 0,1396 |

Затем берутся одноименные месяцы и считается среднее арифметическое их коэффициентов оттока за два года и заполняется Таблица 4.

Считается $TotalChurn(i)$ – сумма коэффициентов оттока для каждого месяца. Для начала зададим начальные значения:

$$TotalChurn(i) = 0, \quad i = январь, февраль, \cdots, декабрь.$$

Обозначив символом " %" оператор остатка от деления, получим:

$$TotalChurn((n+8)\%12+1) = TotalChurn((n+8)\%12+1) + Churn((n+8)\%12+1),$$

$n$= 1, 2, ..., 25.

Цифра "8" в формуле означает, что отсчет месяцев осуществлялся не с января, а с сентября.

1. Определяется к оличество "известных" лет. Так как имеются данные о 24 месяцах, получаем $\frac{24}{12} = 2$ года.
2. Среднее значение коэффициентов:

$$MeanChurn(i) = \frac{TotalChurn(i)}{2}.$$

Среднее арифметическое коэффициентов притока, т.е. $MeanIncome(i)$, считается аналогично.

Таблица 4

**Коэффициенты притока и оттока по месяцам**

| Месяц | Средний коэффициент оттока | Средний коэффициент притока |
|---|---|---|
| Январь | 0,067659 | 0,128679 |
| Февраль | 0,051292 | 0,180598 |
| Март | 0,054044 | 0,195244 |
| … | … | … |
| Октябрь | 0,070714 | 0,147941 |
| Ноябрь | 0,062621 | 0,093500 |
| Декабрь | 0,057185 | 0,080017 |

Для прогнозирования числа учителей на $n$-й месяц получаем формулу, которая значение количества преподавателей на прошлом месяце увеличивает на *MeanIncome* долю, и уменьшает на *MeanChurn* долю:

$$Count(n) = predCount(n-1) - predCount(n-1) \cdot MeanChurn((n+8)\%12+1) + \\ + predCount(n-1) \cdot MeanIncome((n+8)\%12+1).$$

Здесь $predCount(n-1)$ – прогноз количества преподавателей в конце $(n-1)$-го месяца моделью линейной регрессии, $MeanChurn((n+8)\%12+1)$ и $MeanIncome((n+8)\%12+1)$ – соответствующие сезону коэффициенты.

**Оценка модели:**

Точность построенной модели оценим с помощью средней абсолютной ошибки (*MAE*):

$$MAE = \frac{\sum_{i=1}^{n} | trueY_i - predY_i |}{n},$$

где $trueY_i$ – $i$-е реальное значение, $predY_i$ – $i$-е предсказанное значение, $n$ – число испытаний, $i \in 0, \ldots, n$.

Пусть последние пять известных месяцев будут тестовой выборкой. Тогда перестроим модель линейной регрессии и посчитаем коэффициенты оттока и притока с условием, что известно уже не 25 месяцев, а 20. Имеем:

$$MAE = \frac{\sum_{i=1}^{5} | trueY_i - predY_i |}{5} = 62.$$

Для сравнения модель обычной линейной регрессии имеет *MAE*=366. Таким образом, модель, использующая *churn* и *income* коэффициенты, имеет меньшую среднюю абсолютную ошибку, нежели обычная модель линейной регрессии, и увеличивает алгебраический порядок точности с трех до двух.

## 4. ИСПОЛЬЗОВАНИЕ МЕТОДА ТРОЙНОГО ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ

В исследуемом временном ряду наблюдались тренд и ярко выраженная сезонность: значительно возрастало количество уроков в сентябре-октябре, далее плавно спадало. В начале весны наблюдался пик поменьше, а с началом лета количество уроков резко сокращалось. При этом отмечался общий тренд на увеличение количества уроков. В таком формате данных прогнозировалось количество уроков на следующий месяц с помощью тройного экспоненциального сглаживания, или модели Холта-Уинтерса, поскольку легко было выделить 4 сезона и аддитивный тренд [1].

Для предсказания методом тройного экспоненциального сглаживания, или Холта-Уинтерса, использовались следующие формулы [2]:

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \text{ – сглаживание,}$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \text{ – сглаживание тренда,}$$
$$I_t = \beta\frac{\gamma_t}{S_t} + (1-\beta)I_{t-L} \text{ – сезонное сглаживание,}$$
$$F_{t+m} = (S_t + mb_t)I_{t-L+m} \text{ – прогноз,}$$

где $y$ – наблюдение, $S$ – сглаженное наблюдение, $b$ – параметр тренда, $I$ – сезонный параметр, $F$ – прогноз на $m$ периодов вперед, $t$ – индекс, обозначающий период, $\alpha, \beta, \gamma$ – гиперпараметры, которые необходимо подобрать так, чтобы минимизировать значение средней абсолютной ошибки $MAE$, которая рассчитывалась по формуле:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | F_i - y_i |,$$

Число периодов $L$ в данном временном ряду равно 4.

По следующей формуле рассчитывалось начальное значение параметра тренда $b$:

$$b = \frac{1}{L}\left(\frac{y_{L+1} - y_1}{L} + \frac{y_{L+2} - y_2}{L} + \ldots + \frac{y_{L+L} - y_L}{L}\right),$$

затем – начальные значения для сезонных параметров. Для этого находилось среднее по каждому году:

$$A_p = \frac{\sum_{i=1}^{4} y_i}{4}, \quad p = 1, 2,$$

и вычислялись сезонные параметры:

$$I_i = \frac{1}{2}\left(\frac{y_i}{A_i} + \frac{y_{i+12}}{A_2}\right), \quad i = 1 \textit{ означает январь}, i = 2 - \textit{февраль}, \ldots, i = 12 - \textit{декабрь}.$$

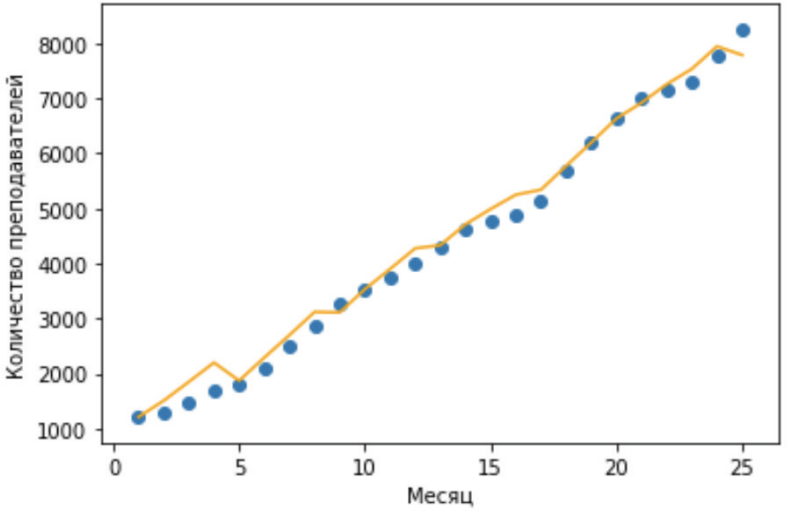


*Рис.2. График тройного экспоненциального сглаживания: желтая кривая – результат сглаживания; синие точки – фактическое число преподавателей в каждом месяце*

После проведения расчетов по приведенным выше формулам и подбора гиперпараметров удалось достичь точности со средней абсолютной ошибкой *MAE*=170 учителей. На рисунке 2 кривая изображает результат сглаживания $S_t$, а точки – фактические значения $y_t$. Для сравнения, модель обычного экспоненциального сглаживания имеет *MAE*=320. Это показывает эффективность учета сезонности и тренда для прогнозирования временных рядов в интересах онлайн-школы.

## 5. ЗАКЛЮЧЕНИЕ

Таким образом, оценивая качество прогнозов всех построенных моделей средней абсолютной ошибкой, получаем, что лучшей моделью является прогноз с использованием коэффициентов оттока и притока. Такая модель позволяет моделировать изменение числа учителей и эффективно прогнозировать необходимое их количество.

***Литература***

1. *Winters P.R.* Forecasting sales by exponentially weighted moving averages // Management Science. – 1960. – Vol. 6. – № 3.
2. *NIST/SEMATECH e-Handbook of Statistical Methods,* http://www.itl.nist.gov/div898/handbook/, 2012
3. *Лукашин Ю.П.* Адаптивные методы краткосрочного прогнозирования временных рядов. – М.: Финансы и статистика, 2003.

# Predicting the Number of Teachers Needed at Online-School

***Gleb B. Sologub****
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000–0002–5657–4826
e-mail: glebsologub@ya.ru

***Vyacheslav A. Pukhov*****
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000–0002–8078–6386
e-mail: csguard26@gmail.com

***Leonid S. Tsyplenkov*******
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000–0001–6010–1223
e-mail: leonidtsyplenkov@gmail.com

The article describes an approach to forecasting the number of teachers to hire by an online English language school, based on an analysis of historical data on the lessons using linear regression and triple exponential smoothing models.

***Keywords:*** time series, time series analysis, time series prediction, online education.

**For citation:**
Sologub G.B., Pukhov V.A., Tsyplenkov L.S. Predicting the Number of Teachers Needed at Online-School. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2020. Vol. 10, no. 2, pp. 39–48. DOI:10.17759/mda.2020100203 (In Russ., abstr. in Engl.).

***References***

1. Winters P.R. Forecasting sales by exponentially weighted moving averages // *Management Science*, 1960. Vol. 6. № 3.
2. *NIST/SEMATECH e-Handbook of Statistical Methods,* http://www.itl.nist.gov/div898/handbook/, 2012
3. Lukashin Yu.P. Adaptivniye metody kratkosrochnogo prognozirovaniya vremennykh ryadov = Adaptive methods for short-term time series forecasting. *Finansy i statistika = Finance and Statistics*, 2003.

***Gleb B. Sologub,** candidate of physical and mathematical sciences, associate professor of the Department of Mathematical Cybernetics, Institute of Information Technologies and Applied Mathematics, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000–0002–5657–4826, e-mail: glebsologub@ya.ru

****Vyacheslav A. Pukhov,** undergraduate student of the Institute of Information Technology and Applied Mathematics, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000–0002–8078–6386, e-mail: csguard26@gmail.com

*****Leonid S. Tsyplenkov,** undergraduate student of the Institute of Information Technology and Applied Mathematics, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000–0001–6010–1223, e-mail: leonidtsyplenkov@gmail.com