Модель точечного прогноза стоимости курсов по английскому языку

в зависимости от определенных факторов

Мария Шустилова

2023

СОДЕРЖАНИЕ

[ВСТУПЛЕНИЕ 2](#_Toc125685828)

[ПАРАМЕТРИЗАЦИЯ 3](#_Toc125685829)

[О НАБОРЕ ДАННЫХ 4](#_Toc125685830)

[ПРЕДВАРИТЕЛЬНАЯ РАБОТА С ДАННЫМИ 4](#_Toc125685831)

[РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ 4](#_Toc125685832)

[МОДЕЛИРОВАНИЕ 5](#_Toc125685833)

[ПРИМЕНЕНИЕ МОДЕЛИ 7](#_Toc125685834)

[ЗАКЛЮЧЕНИЕ + ВЫВОДЫ 8](#_Toc125685835)

# ВСТУПЛЕНИЕ

Задачей данной работы является предсказать стоимость на курсы по английскому в зависимости от нескольких наиболее значимых переменных, которые должны быть либо количественно измеримыми, либо могут быть закодированы в виде дамми переменных (0 или 1)

Изначально планировалось проводить построение модели при задействовании моделей машинного обучения. Однако для того, чтобы модели машинного обучения выдавали корректные прогнозы, требуются десятки тысяч наблюдений. Так как курсы по английскому не исчисляются десятками тысяч, было решено построить намного более простую модель: эконометрическую модель точечного прогноза. Почему не модель интервального прогноза? Во-первых, модель точечного прогноза строится проще. Во-вторых, интервальный эконометрический прогноз может дать слишком широкий интервал, что еще больше запутает человека, устанавливающего цену на курс. Такая модель скорее была бы полезна, если бы вопрос ставился так: «Какая цена точно НЕ может быть установлена за курс?» Ответом было бы все, что находится за пределами выдаваемого моделью интервала. Так что давайте перейдем к построению нашего точечного прогноза.

# ПАРАМЕТРИЗАЦИЯ

Для построения модели зависимости цены курса (в рублях) от других параметров были отобраны следующие количественные характеристики:

1. Длительность курса (в часах). Предполагается, что чем больше длительность курса, тем дороже он стоит. Вопрос заключается в том, на сколько именно дороже
2. Размер скидки на курс (в большинстве случаев этот параметр был равен нулю) (в %). Ожидается, что чем выше скидка, тем выше будет интерес покупателей к курсу
3. Рейтинг онлайн-школы. Предполагается, что чем он выше, тем дороже будут курсы у этой школы

и дамми переменные:

1. Наличие непосредственного общения с учителем в ходе курса (1 – в курс включены живые занятия с учителем, 0 – не включены)
2. Возможность воспользоваться рассрочкой. Предполагается, что курсы, которые можно будет взять в рассрочку, стоят дороже дороже
3. Наличие учебного чата (1 – он есть, 0 – его нет). Гипотеза заключается в том, что наличие чата, в котором ученики могут задавать вопросы, будет в целом положительно влиять на стоимость курса
4. Пожизненный доступ к материалам (1 – предоставляется, 0 – не предоставляется). Эта переменная, скорее всего, будет повышать цену курса
5. Сертификат по окончании обучения (1 – выдается, 0 – не выдается). Люди, вероятно, будут отдавать предпочтение курсам, где в конце обучения выдается сертификат
6. Также была задействована такая характеристика как старт курса: 1 – если курс начинается в любое время, 0 – если курс начинается в определенную дату (например, 30 января). Предполагается, что курсы, которые можно начать проходить когда угодно, более популярны и, соответственно, стоят дороже. Однако это спорный момент и вполне может оказаться в ходе исследования, что эта характеристика работает в другую сторону или же вообще не значима. Но на начальном этапе она была добавлена в датасет на всякий случай

# О НАБОРЕ ДАННЫХ

Набор данных был взят с сайта[[1]](#footnote-1) посредством парсинга. На этом сайте выложена информация по 132 популярным курсам по английскому, и информация постоянно обновляется. Темы курсов довольно разные: как английский для людей разных профессий (фармацевтов, финансистов и т. д.), так и английский для детей или для людей определенного уровня. В датасете представлены курсы от 11 различных онлайн-школ, среди которых преобладает Skyeng (примерно 1/3 от датасета – курсы Skyeng, что не очень хорошо, хотя и не критично)

# ПРЕДВАРИТЕЛЬНАЯ РАБОТА С ДАННЫМИ

1. Закодированы дамми переменные
2. Удалены наблюдения с пропущенными значениями
3. Удалены дублирующие друг друга курсы (очень похожие между собой курсы от одной и той же онлайн-школы и по одной и той же цене)
4. Длительность всех курсов переведена в часы (так как изначально было несколько курсов, чья длительность исчислялась в месяцах; часы, однако, более объективный показатель)
5. Удалены выбросы. Так, в исследовании анализируются курсы, цена которых не превышает 100 тыс рублей, а длительность – 150 часов

После этих операций в датасете осталось 62 наблюдения.

# РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ[[2]](#footnote-2)

Проведена проверка на сбалансированность классов. Так, наблюдается проблема в следующих признаках:

* Рассрочка: всего 3 курса из 62 продается в рассрочку
* Чат: всего в 10 курсах из 62 нет чата

Остальные классы более менее сбалансированы. Но эти два нежелательно использовать в модели, особенно если речь идет о моделях машинного обучения. Возьмем на заметку эти два признака. Возможно, они не будут оказывать статистически значимого влияния на модель и будут удалены.

1. Построена матрица корреляций признаков. Пока что видно, что независимые переменные почти не коррелируют между собой, а на зависимую переменную больше всего положительно влияет наличие на курсе живого общения с учителем (их корреляция составляет 0,5)
2. Отдельно построены распределения стоимости курсов и их продолжительности. Эти распределения напоминают лонгнормальные. Тогда построим отдельно распределения логарифма стоимости курсов и логарифма их продолжительности. В случае продолжительности логарифмирование сработало и ее распределение действительно стало сильно похоже на нормальное; но в случае с ценой логарифмирование лишь немного улучшило ситуацию: получившееся распределение только отдаленно напоминает нормальное. Скорее всего, такая проблема была бы решена наличием б*о*льшего числа наблюдений. Однако из вышесказанного все равно можно сделать вывод, что в исследуемой модели такие параметры как длительность курса и цена должны быть взяты с логарифмом, так как это приближает распределения к нормальным, что уменьшает ошибку прогноза.

# МОДЕЛИРОВАНИЕ

Итак, наша модель выглядит таким образом и содержит 9 независимых параметров:

log(Price) = β0 + βduration\*log(Duration) + βsale\*Sale + βrating\*Rating + βteacher\*Teacher + βcertificate\*Certificate + βlifetime\_access\*Lifetime\_access + βstart\*Start + βchat\*Chat + βinstallment \* Installment

Для оценки коэффициентов при переменных будем использовать R. По результатам оценивания оказалось, что p-value у таких коэффициентов как:

* Скидка
* Пожизненный доступ к материалам курса
* Рейтинг школы
* Старт курса
* Возможность оформления рассрочки
* Выдача сертификата

больше 10%. Это значит, что велика вероятность ошибиться, если принять гипотезу, что эти переменные влияют на модель. Кроме того, p-value для таких параметров как скидка и пожизненный доступ к материалам огромен (более 85%), что говорит о том, что эти параметры не имеют никакого влияния на цену (по крайней мере, на имеющихся данных) или даже ведут к возникновению ошибок в работе модели.

Тогда воспользуемся методом последовательного исключения незначимых параметров, начиная с тех, у которых p-value больше. После исключения по одному каждого из таких параметров только три параметра модели остались значимыми:

* Длительность курса
* Наличие живого общения с учителем
* Наличие чата

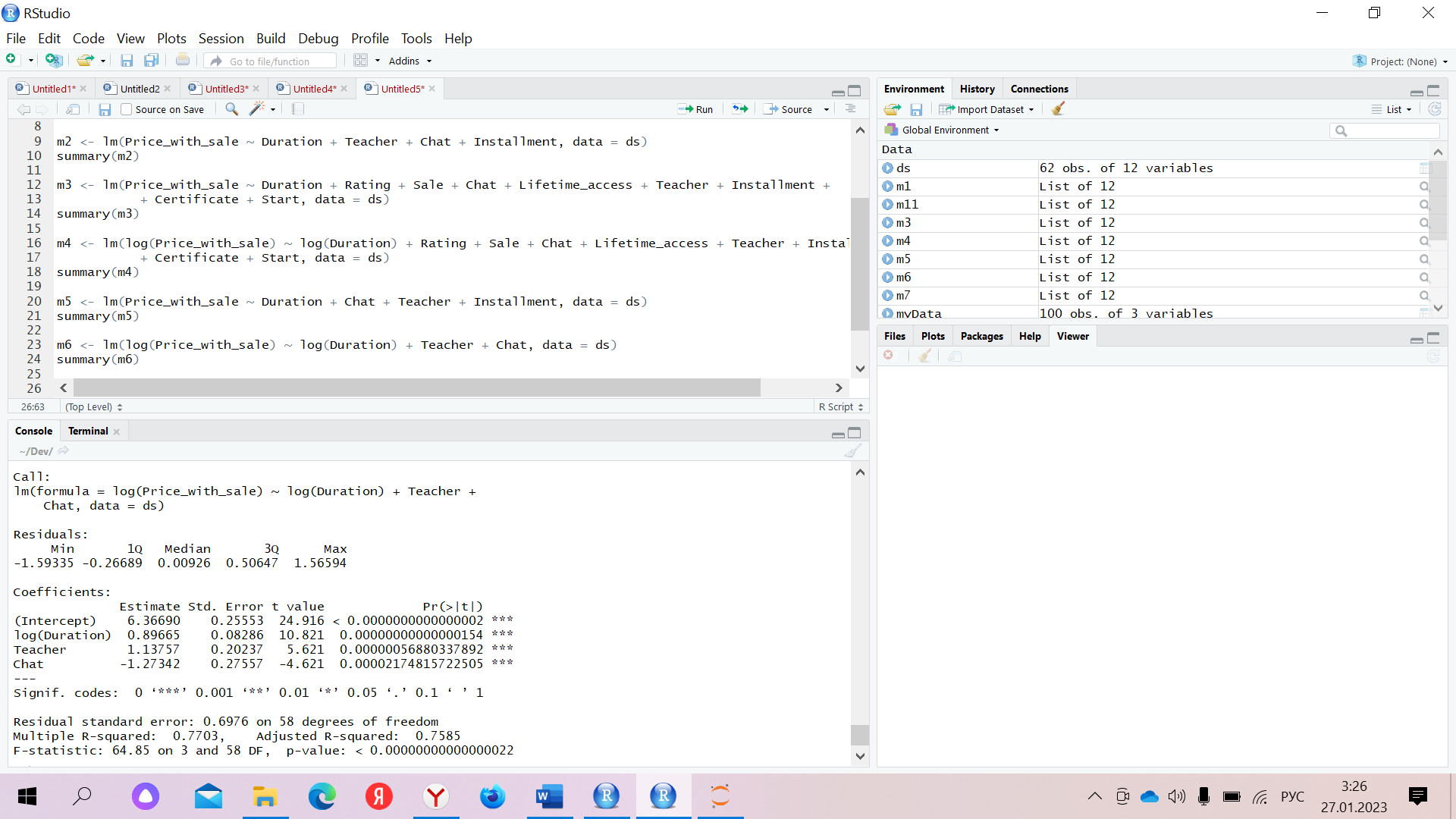


Рисунок 1

Переобозначение переменной «chat»

Наиболее интригующий момент заключается в том, что, исходя из коэффициентов, подобранных R, наличие чата снижает цену на курс. При этом нельзя сказать, что это не валидный результат, так как этот коэффициент значимый (его p-value = 0,002%). То есть при наличии чата закономерно снижение стоимости курса, что противоречит изначальной гипотезе и логике.

Наиболее вероятным объяснением данной ситуации является тот факт, что чат нужен только на курсах, где занятия проходят в группе; более того, чат – чуть ли не обязательный атрибут всех групповых курсов. А значит, он в нашей модели будет индикатором того, что занятия проходят в группе, что делает курс дешевле по сравнению с курсом, где ученик занимается один на один с преподавателем или имеет возможность написать преподавателю в чате с вопросами. Как правило, обозначая в карточке курса наличие чата, онлайн-школы имеют ввиду именно групповой чат, а не диалог с преподавателем (в таком случае пишут обычно, что-то вроде того, что есть возможность общаться с преподавателем в индивидуальном порядке по всем вопросам). Одной из претензий к такому предположению может быть то, что существуют курсы-книги, не включающие ни личного общения с преподавателем, ни работы в группе. В таком случае дамми для чата будет равна 0 и стоимость такого курса вопреки здравому смыслу вырастет. Здесь важно заметить, что, несмотря на такую неточность, стоимость курса в таком случае все равно будет не слишком высокой за счет того, что дамми, отвечающая за наличие учителя, для такого курса будет равна 0, что уменьшит стоимость курса. Кроме того, по английскому сейчас очень мало курсов-книг и в данной подборке таких курсов не более 5%. Из всего вышесказанного следует, что о переменной «chat» можно думать как о переменной «group» (где 1 – это групповые занятия, а 0 – индивидуальное общение с преподавателем)

Но раз в этом плане существуют некоторые неточности, не лучше ли вообще исключить переменную «chat» (или «group») из модели? Нет, потому что модель, построенная с использованием этой переменной, значительно увеличивает прогностическую способность модели: почти на 10%, сокращая вероятность ошибки прогноза. Из-за этого целесообразнее оставить переменную группы. Итоговая модель будет выглядеть следующим образом:

log(Price) = β0 + βduration\*log(Duration) + βteacher\*Teacher + βchat\*Group

С оцененными коэффициентами:

log(Price) = 6.36690 + 0.89665\*log(Duration) + 1.13757\*Teacher - 1.27342\*Group

Таким образом, стоимость курса может быть предсказана с помощью всего лишь трех параметров.

Что примечательно, после удаления «лишних» переменных прогностическая способность модели (R2 adjusted) даже немного увеличилась и составила 76%

# ПРИМЕНЕНИЕ МОДЕЛИ

И, наконец, для применения полученной модели требуется:

1. Задать эти три параметра:

* Продолжительность курса (ч)
* Наличие личного взаимодействия с преподавателем (1 – есть, 0 – нет)
* Занятия проходят в группе или нет (1 – да, 0 – нет)

1. Зайти в любой Python-компилятор. Например, сюда: <https://www.onlinegdb.com/online_python_compiler>
2. Вместо текста, написанного в окошке компилятора, ввести следующий код, сохраняя отступы:

import math

duration = float(input('Введите количество часов: '))

teacher = float(input('Учитель (1 или 0): '))

group = int(input('Группа (1 или 0): '))

ln\_duration = math.log(duration)

answer = round(math.e\*\*(6.36690 + 0.89665 \* ln\_duration + 1.13757 \* teacher - 1.27342 \* group), 0)

print(f'Цена на курс: {answer}')

# ЗАКЛЮЧЕНИЕ + ВЫВОДЫ

Получившаяся модель предсказывает цену на курс по английскому (в рублях) с точностью около 76%, которая является приемлемой для эконометрических проектов. Однако со временем точность будет снижаться, что потребует обновления данных и, возможно, даже пересборки модели. В случае, если вы хотите воспользоваться моделью, а со времени ее создания (середина января 2023 года) прошло более 1-2 лет, рекомендую индексировать полученный результат на размер инфляции за прошедшие годы. Но это не панацея, так как в течение нескольких лет может меняться влияние одного или нескольких из наших трех факторов на цену, а также могут появляться новые значимые факторы. Однако если таких изменений явно не наблюдается, эта модель может помочь при оценивании реальной стоимости курса.

В то же время, получаемое на выходе число служит лишь ориентиром по стоимости, так как данное исследование замеряет лишь количественные характеристики. При этом качественные характеристики не замеряются и не могут быть точно замерены. Таким образом, в рамках данного исследования не могли быть проанализированы такие характеристики как знания и навыки преподавателей, дизайн интернет-страницы школы, качество ее маркетинга и т. д. и если в Вашем курсе наблюдается существенное количество дополнительных бонусов, которые невозможно измерить количественно, то, вероятно, ваш курс стоит дороже той суммы, которую предсказывает модель. Из-за этого рекомендуется использовать эту модель для оценки стоимости стандартных курсов, а не их премиум-собратов.

Стоит добавить, что в ходе исследования выяснилось, что на стоимость курса не влияют скидка на него и возможность получить рассрочку, а также сертификат по окончании курса.

1. <https://www.sravni.ru/kursy/anglijskij-yazyk/region/moskva/> [↑](#footnote-ref-1)
2. подробности см. в коде [↑](#footnote-ref-2)