Автоматическое выявление наиболее выгодных автомобильных лотов на торгах по банкротству

Жучкова Светлана

21.03.2019

Введение

Торги, или аукционы, по банкротству – предусмотренная законодательством России процедура возмещения средств кредитору от должника, признанного банкротом, через продажу имущества последнего. Торги организуются на специальных открытых электронных площадках и состоят из нескольких этапов, на последнем из которых имущество реализуется, как правило, по сильно сниженной цене. Легальность торгов, открытый доступ к участию в них и возможность покупки имущества по цене ниже рыночной объясняют интерес к таким торгам со стороны их участников. С 2014 года наблюдается существенный рост числа соответствующих поисковых запросов[1] на территории России в поисковой системе Google[2] (см. Рис. 1 [Google Trends]), а само число участников российских торгов на 2018 год увеличилось почти вдвое по сравнению с 2014 годом [Мальцев, 2018]. Увеличиваются и объёмы торгов: в 2018 году наблюдался

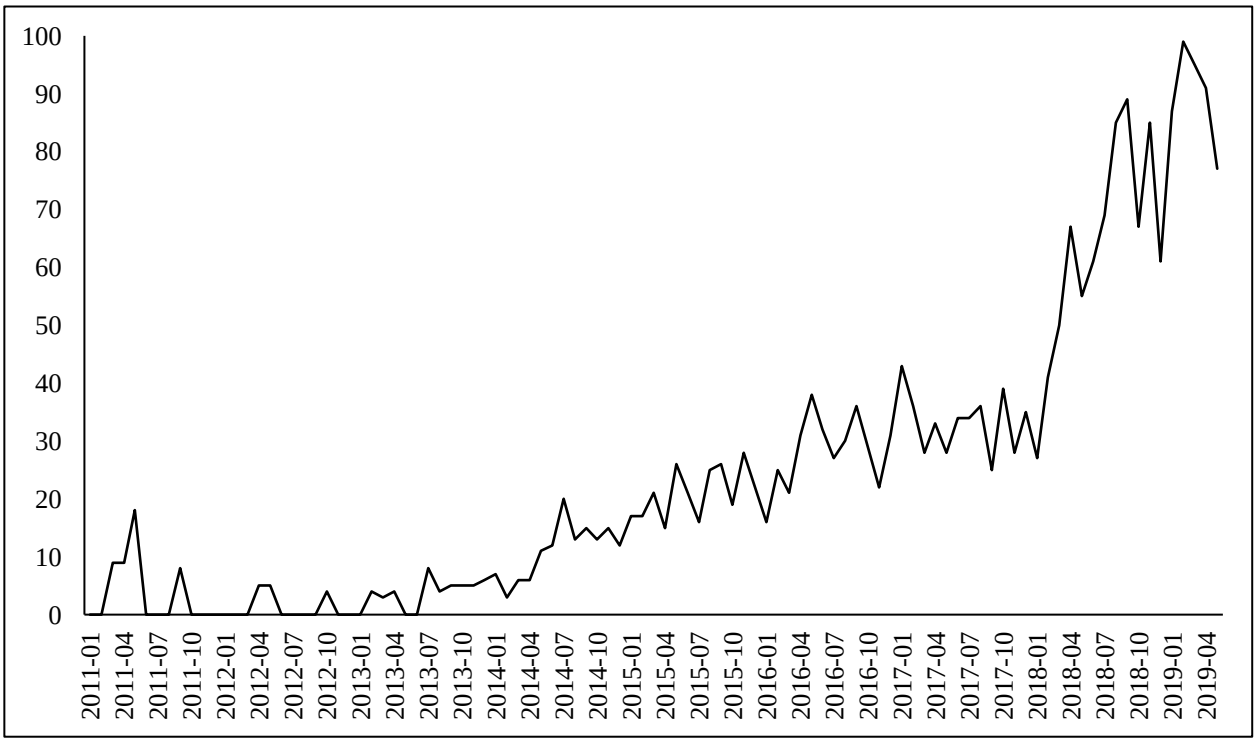

Рис. 1 Динамика поискового запроса "торги по банкротству" в России в поисковой системе Google (2011-2019)

[1] Поскольку торги организуются исключительно на онлайн-площадках, мы считаем поисковые запросы одним из индикаторов проявления интереса к ним.

[2] На графике представлена внутренняя метрика Google-поиска interest over time, показывающая долю поисковых запросов, приходящуюся на конкретный период времени, по отношению к максимальному числу запросов в том же регионе за весь запрашиваемый период времени. Таким образом, максимальное значение метрики приходится на февраль 2019 (100), когда наблюдалось самое большое число соответствующих запросов, а все остальные точки на графике построены по отношению к этой точке.

заметный рост как выставленных, так и реализованных лотов на электронных торговых площадках: на 46% и 65% по сравнению с 2017 годом соответственно [там же].

Участие в торгах по банкротству через покупку и последующую перепродажу имущества становится популярным способом заработка. Тем не менее, этот способ сопряжён с определёнными трудностями: большое количество площадок[3] и лотов на них, отсутствие структурированной информации о лотах, необходимость обладать экспертным знанием о рыночных стоимостях лотов – эти факторы могут приводить к неоптимальному выбору лота и потере потенциальной прибыли, если в дальнейшем предполагается перепродажа имущества. Несмотря на то что существует большое количество образовательных материалов (очных и заочных курсов, пособий, обучающих видео) по участию в торгах, тем не менее, итоговый выбор лота во многом зависит от «человеческого фактора», способности самого участника торгов выбрать среди многих вариантов наиболее выгодный.

Настоящая работа носит прикладной характер и посвящена разработке алгоритма для *автоматического* выявления наиболее выгодных автомобильных лотов на торгах по банкротству. Автомобильные лоты выбраны как стартовая точка для будущего приложения, способного работать с любыми лотами, поскольку данные по рыночной стоимости именно автомобилей являются наиболее доступными среди всех остальных категорий лотов. Ранжирование автомобилей по «выгодности» происходит на основе двух критериев: относительной маржи и вероятности продать машину за срок не более одного месяца. Для решения задачи такого ранжирования среди прочих отобраны две модели: градиентный бустинг, предсказывающий рыночную стоимость автомобиля с торгов, и логистическая регрессия, предсказывающая вероятность реализовать автомобиль за указанный срок. Данные, используемые для разработки моделей, собраны с помощью процедуры автоматического извлечения данных (веб-скрапинга) с сервисов TBankrot [Электронные торги по банкротству] и Auto.Ru [Авто.ру]. Результатом работы на текущем этапе стала апробация построенных моделей и формирование предварительного списка наиболее выгодных автомобильных лотов на данных января 2019 года.

Процедура организации торгов по банкротству

Процедура проведения торгов по банкротству в России регулируется федеральным законом от 26.10.2002 N 127-ФЗ «О несостоятельности (банкротстве)». Согласно ему,

[3] Так, только на официальном федеральном ресурсе, агрегирующем информацию о должниках и торгах по банкротству, зарегистрированы около 50 электронных торговых площадок [Единый федеральный реестр].

должниками признаются «гражданин … или юридическое лицо, оказавшиеся неспособными удовлетворить требования кредиторов по денежным обязательствам» в течение определённого срока (три месяца). Реализация имущества гражданина – одна из реабилитационных процедур, применяемых к банкроту «в целях соразмерного удовлетворения требований кредиторов», которая организуется, в общем случае, в виде аукциона на специальных электронных торговых площадках. К имуществу, реализуемому на торгах, относится «недвижимое имущество, ценные бумаги, имущественные права, заложенное имущество, предметы, имеющие историческую или художественную ценность» и др. [Федеральный закон от 26.10.2002 N 127-ФЗ].

Торги по банкротству могут включать в себя до трёх этапов в следующем порядке: аукцион, повторные торги и публичное предложение. Аукцион (также называется первыми торгами) – первый этап процесса реализации имущества, в ходе которого участники торгов делают ставки на *повышение* стоимости лота. Шаг аукциона – величина повышения стоимости – обычно устанавливается правилами заранее. Стартовая цена имущества, выставленного на продажу, определяется в ходе процедуры оценки имущества должника. Повторные торги – второй этап процесса реализации имущества, который назначается в случае, если лот не был выкуплен на первом этапе. На этом этапе стартовая цена лота понижается на 10% по сравнению с той, которая была выставлена на предыдущем этапе. Участники так же делают ставки на *повышение* стоимости лота. И на первом, и на втором этапе победителем становится тот участник, который предложил наибольшую цену за лот. Наконец, последний этап – публичное предложение – организуется, если лот не был выкуплен и на втором этапе. Стартовая цена лота выставляется на прежнем уровне, однако теперь она *уменьшается* через определённые промежутки времени. Как правило, победителем торгов на этом этапе становится участник, подавший заявку на покупку лота раньше других.

Считается, что даже на первый этап торгов предложения поступают уже с ценами ниже рыночных [Аукционы и торги по банкротству в 2019 году…]. Успешное участие в торгах в связи с этим требует либо умения своевременно выявить выгодный лот на одном из первых двух этапов, либо возможности первым подать заявку на лот, находящийся на последнем этапе – обе задачи можно считать предпосылками создания инструмента для автоматического выявления наиболее выгодных лотов, чему и посвящена настоящая работа.

Методология исследования

На текущем этапе основной целью стала разработка моделей, в результате применения которых можно получить предварительный список наиболее выгодных автомобильных лотов на основе данных, собранных в январе 2019 года. Масштабирование моделей на более актуальные данные и на другие категории лотов, а также оформление разработанных алгоритмов в форму веб-приложения или бота запланированы на будущие этапы работы.

Проведённую работу по созданию моделей можно разделить на следующие основные этапы (они же задачи исследования):

1. Выбор источника данных о лотах, представленных на торгах, и сбор данных из этого источника (далее – база 1).
2. Выбор источника данных о рыночной стоимости автомобилей и сбор данных из этого источника (далее – база 2).
3. Предобработка баз 1 и 2, отбор признаков для дальнейшего использования в моделировании.
4. Построение и сравнение моделей, предсказывающих рыночную стоимость автомобилей из базы 1, выбор наиболее точной модели (далее – модель 1).
5. Сбор дополнительных данных о сроке продажи автомобилей.
6. Построение и сравнение моделей, предсказывающих срок продажи автомобиля, выбор наиболее точной модели (далее – модель 2).
7. Ранжирование автомобилей по предсказанным рыночным стоимостям и сроку продажи, определение наиболее выгодных лотов.

Среди всех перечисленных задач наиболее важной содержательно можно признать задачу 4 – предсказание рыночной стоимости автомобиля. От точности прогнозов именно этой модели зависит потенциальная прибыль участника торгов, который ориентируется на перепродажу автомобиля, или реальная выгодность покупки для участника, не заинтересованного в перепродаже, а совершающего покупку «для себя».

Задача моделирования стоимости поддержанного автомобиля не нова, но до сих пор актуальна: заметное количество публикаций по этой теме начинает появляться в конце 2000-х годов [Oprea, 2010; Richardson, 2009; Soejima, Hirose, 2011; Wu et al., 2009], и вплоть до текущего времени аналогичные работы регулярно презентуются на конференциях [Chen et al., 2017; Sun et al., 2017]. Большинство из этих работ носят сравнительный характер: в них сравниваются либо различные методы машинного обучения [Gegic et al., 2019; Noor, Jan, 2017; Pudaruth, 2014], либо методы из разных подходов – больше свойственных эконометрике или машинному обучению [Chen et al.,

2017; Ozgur et al., 2016]. Набирает популярность и использование нейронных сетей для оценки стоимости машины [Sun et al., 2017; Shen et al., 2011]. Согласно результатам существующих исследований, набольшую точность и устойчивость в случае наличия множества признаков показывают ансамблевые методы машинного обучения (особенно часто в работах фигурирует случайный лес [Pal et. al, 2018]), а в случае, когда признаков немного – принципиальной разницы между ансамблевыми методами и традиционными эконометрическими (линейной регрессией) нет [Chen et al., 2017]. В своём исследовании при построении модели предсказания стоимости автомобиля мы также сравниваем различные методы, свойственные разным подходам: линейную регрессию, больше характерную для эконометрического анализа, и ансамблевые методы, относящиеся к машинному обучению. Мы не прибегаем к использованию нейронных сетей, поскольку, как будет описано далее, работаем с небольшим числом признаков, а нейронные сети выглядят в этом случае неуместно сложным алгоритмом.

Результаты исследования

Представим результаты исследования поэтапно – согласно шагам, определённым в методологии исследования.

Этап 1: выбор источника данных о лотах, представленных на торгах, и сбор данных из этого источника (далее – база 1)

Как уже было отмечено ранее, количество электронных торговых площадок, размещающих лоты для торгов по банкротству, достаточно велико: только на официальном федеральном ресурсе, содержащем информацию о должниках и торгах по банкротству, зарегистрированы 46 таких площадок [Единый федеральный реестр]. Тем не менее, существуют сайты-агрегаторы, объединяющие информацию о лотах сразу со многих площадок. Один из таких сайтов – TBankrot.ru – стал источником данных о лотах в нашей работе. Сайт был выбран по нескольким причинам: во-первых, на момент сбора данных – январь 2019 года – этот сайт агрегировал информацию о наибольшем числе лотов (более одного миллиона) среди всех других сайтов-агрегаторов; во-вторых, сайт содержит рубрикатор, который даёт возможность ограничиться лотами, удовлетворяющими определённым критериям; в-третьих, сайт обладает простой html-структурой, что позволяет без трудностей собирать с него информацию посредством веб-скрапинга – автоматического извлечения данных с веб-страниц.

Для сбора данных были отобраны лоты с уже завершённых торгов категории «транспорт и техника», выставленные на продажу в Москве и Московской области (каждый из этих критериев задавался с помощью упомянутого рубрикатора). По каждому

из лотов извлекалась следующая информация (при наличии): идентификатор торгов, название электронной торговой площадки, идентификатор лота, факт наличия фотографий лота в описании, текстовое описание лота, стартовая цена, этап аукциона, шаг аукциона, задаток, идентификатор должника (ФИО или название юридического лица), ФИО победителя аукциона, дата получения сообщения от победителя, время получения сообщения от победителя, дата публикации лота, время публикации лота, дата проведения торгов, итоговая цена, отношение итоговой цены к стартовой (%). Объём наблюдений в собранной базе составил 4765 лотов. В дальнейшем анализе на текущем этапе участвовала переменная с текстовым описанием лотов – именно она содержит основную информацию о продаваемом имуществе – и переменная со стартовой ценой. Остальные переменные, тем не менее, могут быть полезны для более детального описания процедуры проведения торгов за рамками настоящей работы.

Этап 2: выбор источника данных о рыночной стоимости автомобилей и сбор данных из этого источника (далее – база 2)

В качестве источника данных о рыночной стоимости автомобилей была выбрана наиболее крупная площадка с объявлениями о продаже автомобилей в России – Auto.Ru – содержащая объявления как о поддержанных автомобилях, размещаемых частными лицами, так и новых автомобилях без пробега, продаваемых автомобильными салонами. Данные были также собраны с помощью процедуры веб-скрапинга, однако, в отличие от сайта TBankrot, с этого сайта данные извлекались не напрямую из html-кода, а через извлечения json-файлов – фрагментов внутренней базы данных описываемой площадки. Данные были также ограничены объявлениями из Москвы и Московской области. Поскольку сайт накладывает ограничения на выдачу результатов поиска нужных объявлений, данные собирались отдельно для каждой марки автомобиля, представленной на сайте (но не более 3663 объявлений), а затем агрегировались в единую базу. Таким образом было получено 83701 наблюдение.

Внутренняя база сайта Auto.Ru содержит большое количество признаков – 302 переменные, описывающие характеристики автомобиля (в том числе подробно технические характеристики), продавца и самого объявления. Однако для цели исследования необходимо, чтобы признаки из этой базы – те, на которых будут обучаться модели предсказания рыночной цены и скорости продажи – были сопоставимы с признаками, находящимися в базе 1. Этому сопоставлению посвящён следующий этап исследования.

Этап 3: предобработка баз 1 и 2, отбор признаков для дальнейшего использования в моделировании

Как уже было упомянуто, в базе 1 фактически только одна из переменных содержит информацию об автомобиле, выставленном на лот – это его текстовое описание. Единой структуры этого описания не существует – каждый автор объявления указывает ту информацию, которую считает нужной. Кроме того, эта информация крайне «зашумлённая»: например, авторы описаний могут указывать названия автомобилей на разных языках, использовать специальные аббревиатуры, допускать опечатки. Эти обстоятельства сильно ограничили дальнейший анализ: необходимо было вычленить из таких текстов максимальное количество признаков, имеющих аналоги в базе 2.

Для задачи извлечения признаков мы обратились к помощи асессоров, и процесс разметки базы строился следующим образом: около 50 описаний были детально проанализированы и размечены нами, в ходе чего были определены те признаки из базы 2, которые чаще всего встречаются в описаниях из базы 1, были приведены к единому виду возможные категории этих признаков, и на основе этого фрагмента базы асессорами была размечена оставшаяся часть наблюдений. Таким образом из текстовых описаний были извлечены следующие признаки: релевантность лота[4], марка автомобиля (на английском языке), модель автомобиля (на английском языке), марка автомобиля (на русском языке), модель автомобиля (на русском языке), дополнительные характеристики модели, год выпуска, пробег (в км), идентификационный номер (VIN), наличие ареста на автомобиль, нахождение автомобиля в залоге, наличие ограничений на регистрацию автомобиля, факт упоминания повреждений, код цвета автомобиля, название цвета автомобиля, тип должника (физическое или юридическое лицо), факт указания номер паспорта автомобиля, мощность автомобиля (в лошадиных силах), объём двигателя (в литрах). Несмотря на возможность извлечь эти признаки, большинство описаний не содержало многих из этих характеристик (информация о пропущенных значениях представлена в Таблице 1). Кроме того, после отбора только релевантных лотов количество наблюдений сократилось до 2225. В связи с этим на текущем этапе мы решили использовать признаки с наименьшим количеством пропусков: марка и модель автомобиля, год его выпуска. Наблюдения с пропущенными значениями по этим переменным были также исключены.

[4] Поскольку на сайте TBankrot автомобильные лоты входят в более общую категорию «транспорт и техника», в базу данных по лотам попали также и прочие транспортные средства: мотоциклы, грузовые автомобили и т.д. Также в базу ошибочно попали лоты, реализуемые за пределами Москвы и Московской области. Для дальнейшего анализа были отобраны только легковые автомобили, находящиеся именно в двух указанных регионах.

Таким образом, база, которая в дальнейшем используется именно для предсказания цены лота (не для обучения моделей) и определения наиболее выгодных из предложений, была сокращена до 1990 наблюдений.

Таблица 1

Распределение пропущенных значений по переменным из базы 1

Переменная	Описание	Процент пропусков
id	Идентификатор	0%
lot_info	Текстовое описание	0%
is_relevant	Релевантность	0%
mark_en	Марка (на английском)	2%
mark_ru	Марка (на русском)	2%
model_en	Модель (на английском)	3%
model_ru	Модель (на русском)	3%
state_not_beaten	Небитое состояние	6%
year	Год выпуска	11%
vin	VIN	20%
color_hex	Код цвета	47%
color_human_name	Название цвета	47%
power	Мощность (лс)	59%
mileage	Пробег	72%
model_add	Модель (доп. информация)	83%
debtor	Тип должника	91%
encumbrance_deposit	Нахождение в залоге	91%
encumbrance_arest	Наличие ареста	95%
pts	Факт указания ПТС	96%
encumbrance_register	Наличие ограничения на регистрацию	97%
defects	Наличие дефектов	98%
volume	Объём двигателя (л)	98%

Согласно экспертному мнению, полученному нами, и существующим подходам к моделированию стоимости автомобиля, трёх оставшихся переменных содержательно недостаточно для построения качественной модели: необходимо также учитывать как минимум пробег автомобиля и объём его двигателя. Пробег автомобиля указывается в небольшом количестве описаний лотов с торгов (28%), поэтому от использования этой переменной мы отказались. Кроме того, пробег автомобиля довольно тесно коррелирует с годом выпуска (в базе 2 значение выборочной корреляции этих переменных находится на уровне -0,5), информация о котором есть в большинстве описаний лотов. Объём двигателя как таковой также встречается в текстах описаний крайне редко (2%) – в случае, если мы попробовали бы вычленить такую переменную и использовать её при моделировании и

дальнейшем отборе машин, мы бы лишились большей части потенциальных автомобилей – «кандидатов» на покупку. Повторно обратившись к экспертам, мы получили следующий совет: у более дорогих автомобилей больше вариантов возможных объёмов двигателей. Это означает, что в качестве дополнительной переменной можно использовать не объём двигателя как таковой, а количество возможных вариантов объёма двигателя в разрезе различных марок и моделей машин. Эта переменная была получена из базы 2: для каждого сочетания марки и модели машины мы рассчитали количество уникальных объёмов двигателя в этой базе для данного сочетания, предполагая, что большой объём этой базы позволяет получить достаточно точные оценки количества вариантов объёма двигателя у автомобиля. Фактически новая переменная представляет из себя эффект взаимодействия марки и модели машины, однако в отличие от традиционного взаимодействия, значения в новой переменной получились количественными, а не категориальными (это некий «оцифрованный» эффект взаимодействия). Это обстоятельство позволяет в дальнейшем использовать новую переменную без дополнительных преобразований.

На рисунках ниже представлены наиболее часто встречающиеся марки машин из обеих баз, описательная статистика по году выпуска автомобилей из обеих баз, а также информация о количестве вариантов объёма двигателя для разных сочетаний марок и моделей машин – единая для двух баз.

Топ-10 наиболее часто встречающихся марок в обеих базах совпали, однако в базе 2 доля каждой из марок не может превышать 4,4% – это связано с упомянутыми выше ограничениями парсинга этого сайта: по любой из марок машин можно извлечь не больше 3663 объявлений, как бы много на сайте их ни было.

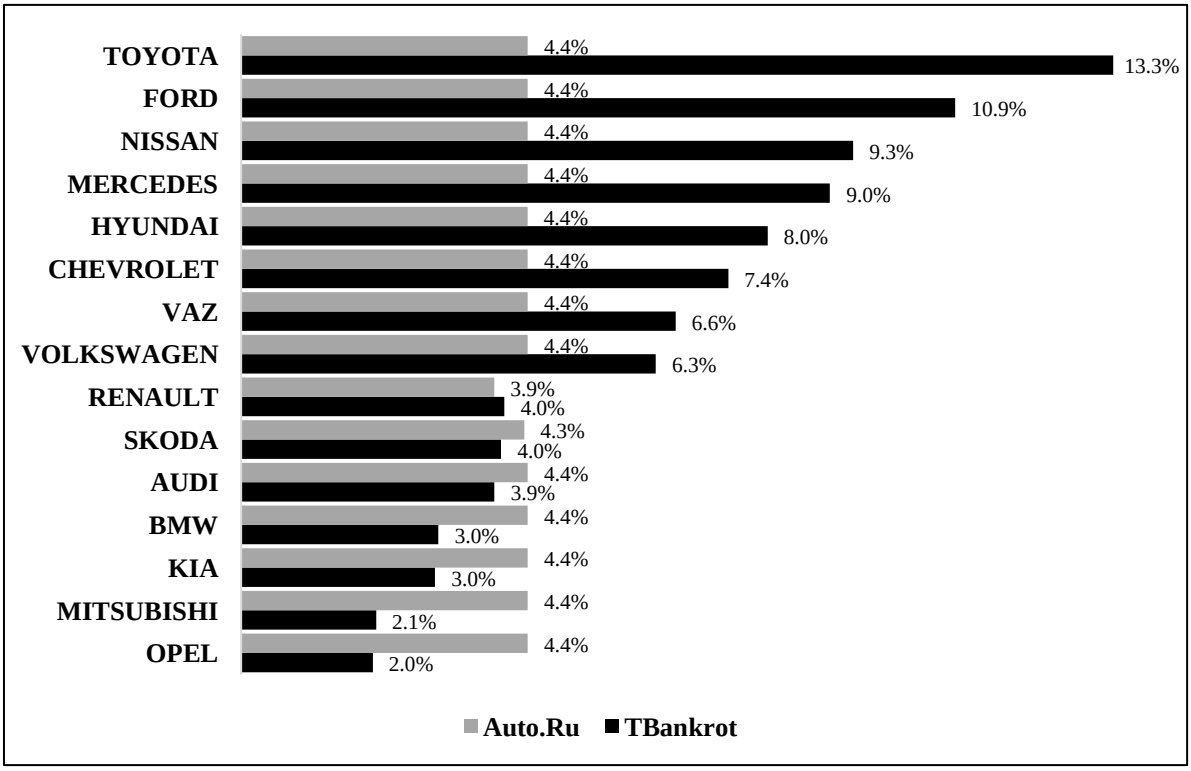

Рис. 2 Топ-10 наиболее часто встречающихся марок автомобилей в двух базах

База Auto.Ru содержит более старые (вплоть до 1935 г.) и более новые (до 2019 г.) машины, чем база TBankrot, однако все меры центральной тенденции в двух базах совпадают (см. Таблицу 2).

Таблица 2

Описательная статистика по году выпуска машин из обеих баз

Мера	TBankrot	Auto.Ru
Минимум	1978	1935
Максимум	2017	2019
Медиана	2011	2011
Среднее	2009	2009
Мода	2012	2012
N	1990	83701

На рис. 3 представлены модели машин с наибольшим числом вариантов объёма двигателя. Хотя представленные на гистограмме данные в некоторой мере подтверждают предположение о том, что более дорогие машины имеют большую вариацию объёма двигателя, статистически эта гипотеза не подтвердилась: ранговый коэффициент корреляции между числом вариантов объёма двигателя и средней ценой автомобиля с таким сочетанием марки и модели составил 0,06. Тем не менее, мы оставили эту переменную для моделирования как эффект взаимодействия марки и модели.

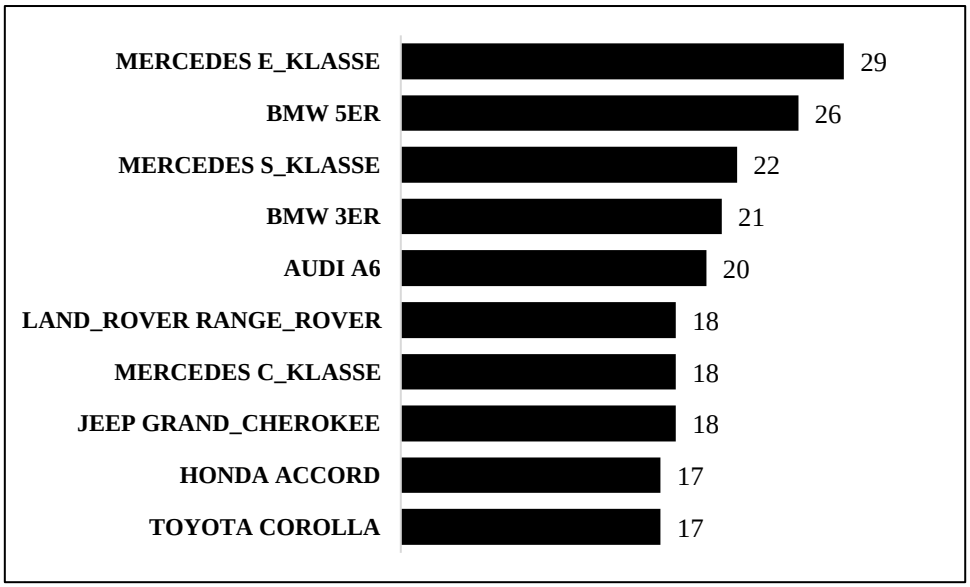

Рис. 3 Модели машин с наибольшим числом вариантов объёма двигателя

Этап 4: построение и сравнение моделей, предсказывающих рыночную стоимость автомобилей из базы 1, выбор наиболее точной модели (далее – модель 1)

Итак, для построения модели по предсказанию рыночной стоимости автомобиля были использованы 83701 наблюдение из базы 2 и пять признаков: марка машины, модель

машины, год выпуска, количество вариантов объёма двигателя для данного сочетания марки и модели и цена автомобиля, заявленная в объявлении. Целевой переменной стала цена автомобиля, а целевой метрикой – коэффициент детерминации R_2. Выборка была поделена на две части: 67% на обучающую и 33% на тестовую.

Среди моделей сравнивались линейная регрессия без регуляризаторов, линейная регрессия с L1- и L2-регуляризаторами, бэггинг на решающих деревьях, случайный лес и градиентный бустинг на решающих деревьях. Таким образом, сравнивались модели из разных подходов: эконометрического и машинного обучения. Для линейной регрессии категориальные переменные (марка и модель машины) были дихотомизированы, для остальных методов – использовались без преобразований.

Для подбора гиперпараметров в тех моделях, где это необходимо (ансамблевых), мы использовали процедуру автоматического определения наилучшей модели по целевой метрике через кросс-валидацию: при каждом новом сочетании параметров исходная обучающая выборка делилась на три подвыборки, каждая из которых поочерёдно удалялась и использовалась как тестовая выборка для расчёта целевой метрики, а модель обучалась на оставшихся данных. Лучшей считалась та модель, которая имела наивысшее среднее значение целевой метрики по этим трём подвыборкам. В рамках каждого из шести методов мы обучили модель с выбранными гиперпараметрами на всей обучающей выборке, рассчитали значение на общей тестовой выборке и выбрали ту из моделей, которая оказалась наименее переобученной. Итоговой моделью стал градиентный бустинг с максимальной глубиной 5, количеством деревьев 500, минимальным числом наблюдений в узле 1 и среднеквадратичной ошибкой в качестве функции потерь. Для этой модели R_2 на обучающей выборке составил 0,93, на тестовой – 0,88. Заметим, что такие же значения точности модели были получены в прочих исследованиях со сравнением моделей для предсказания стоимости автомобиля, использующих при этом гораздо больше признаков при обучении [Pal et al., 2018].

Согласно полученной модели, наиболее важным признаком оказалась модель автомобиля, затем в порядке убывания важности следуют год выпуска, марка и количество вариантов объёма двигателя.

Этап 5: сбор дополнительных данных о сроке продажи автомобилей

Отдельной задачей для определения наиболее выгодных автомобилей мы поставили предсказание срока продажи автомобиля, купленного на торгах – в случае, если идёт речь о последующей перепродаже автомобиля и получении прибыли. Хотя соответствующих данных в исходной базе 2 не было, на самом сайте Auto.Ru при продаже

автомобиля на странице с объявлением выставляется отметка, что автомобиль продан и объявление не актуально (после чего объявление уже не доступно из общего поиска, но доступно по прямой ссылке). В связи с этим мы предприняли следующее: для случайной подвыборки объявлений из базы 2 осуществлять ежедневный мониторинг страниц по прямым ссылкам на них и проставлять в базу дату, когда на странице появилась отметка о продаже автомобиля. После трёхмесячного мониторинга 93% автомобилей из подвыборки были проданы; затем было рассчитано количество дней, прошедших с даты публикации объявления до даты продажи автомобиля. Тем автомобилям, которые не продались за период проведения мониторинга, было выставлено значение 550 дней, приблизительно соответствующее сроку в полтора года. Таким образом для 3762 автомобилей появились сведения о сроке их продажи.

Этап 6: построение и сравнение моделей, предсказывающих срок продажи автомобиля, выбор наиболее точной модели (далее – модель 2)

Те же признаки, которые участвовали в моделировании стоимости автомобиля, стали основой и для моделирования срока продажи автомобиля. Целевой переменной исходно было количество дней, прошедших с даты публикации объявления до даты продажи автомобиля – количественная переменная, полученная на предыдущем этапе. Однако ни один из сравниваемых алгоритмов (их состав и принцип сравнения были теми же, что и на этапе 4) не показал приемлемого качества: для всех моделей R_2 на тестовой выборке принимал отрицательные значения (на обучающей – варьировался от 0,2 до 0,56).

В связи с этим мы решили отказаться от использования исходной количественной переменной и перейти от задачи регрессии к задаче классификации – предсказывать не точное количество дней до продажи автомобиля, а факт продажи автомобиля за определённый срок. Граница этого срока также была согласована с экспертами, заинтересованными в использовании разрабатываемой модели, и составила один месяц (30 дней). Таким образом, целевой переменной теперь стала дихотомическая переменная о факте продажи автомобиля за месяц, а целевой метрикой – средняя F-метрика по обеим категориям зависимой переменной для тестовой выборки. Алгоритмами для сравнения стали методы бинарной классификации: логистическая регрессия, дерево решений, бэггинг и градиентный бустинг на деревьях решений, а также случайный лес. Так же, как и на этапе 4, после подбора оптимальных гиперпараметров через кросс-валидацию лучшие модели в рамках каждого метода сравнивались по критерию непереобученности – и в этот раз наименее переобученной оказалась логистическая регрессия. При этом качество всех сравниваемых на этом этапе моделей было невелико: средняя F-метрика на

тестовой выборке находилась на уровне 0,54-0,57. Несмотря на такие результаты, мы решили воспользоваться построенной моделью, но со следующими оговорками: поскольку ошибочная классификация больше свойственна для тех объектов, которые по итогам использования модели имеют «пограничные» вероятности обоих исходов (около 0,5), для последующего ранжирования мы отбираем только те автомобили, которым предсказана именно высокая вероятность быть проданными в течение одного месяца (в которых алгоритм «уверен» больше всего). В качестве порога такой «высокой вероятности» было взято значение 0,7.

Этап 7: ранжирование автомобилей по предсказанным рыночным стоимостям и сроку продажи, определение наиболее выгодных лотов

После того как обе модели были обучены на обучающей выборке из базы 2, проверены на тестовой выборке из этой же базы, мы перешли к их применению на базе 1 и отбору наиболее выгодных лотов. В общем виде алгоритм отбора был следующим:

1. Предсказание рыночной стоимости автомобиля с торгов на основе его марки, модели, года выпуска и количества вариантов объёма двигателя для данного сочетания марки и модели – с помощью модели 1.
2. Расчёт абсолютной и относительной маржи – разницы между предсказанной рыночной стоимостью и стоимостью на торгах и отношения этой разницы к стоимости на торгах (в процентах) соответственно.
3. Предсказание вероятности продать автомобиль за срок не более одного месяца на основе его марки, модели, года выпуска и количества вариантов объёма двигателя для данного сочетания марки и модели – с помощью модели 2.
4. Отбор только тех автомобилей, для которых вероятность их продажи за срок не более одного месяца составляет 0,7 и выше.
5. Ранжирование отобранных автомобилей по относительной марже. Автомобили, попавшие в этот список, упорядочены по «выгодности» покупки с целью последующей перепродажи.

Фрагмент получившегося списка автомобилей (первые 10 наиболее выгодных вариантов) представлен в Таблице 4.

Таблица 4

Топ-10 наиболее выгодных для покупки автомобилей на торгах (январь 2019)

Автомобиль	Год выпуска	Цена на торгах (руб.)	Предсказанная цена (руб.)	Маржа (руб.)	Маржа (%)	Вероятность продажи за месяц
RENAULT DUSTER	2013	250000	528934	278934	111,6	0,86

Автомобиль	Год выпуска	Цена на торгах (руб.)	Предсказанная цена (руб.)	Маржа (руб.)	Маржа (%)	Вероятность продажи за месяц
HYUNDAI SOLARIS	2015	300000	547760	247760	82,6	0,75
RENAULT LOGAN	2011	161000	290609	129609	80,5	0,80
CHERY TIGGO	2013	283985	458140	174155	61,3	0,77
TOYOTA COROLLA	2012	439994	664503	224509	51,0	0,74
HYUNDAI SOLARIS	2007	203400	302877	99477	48,9	0,75
MITSUBISHI LANCER	2009	256230	381400	125170	48,9	0,73
SSANG_YONG ACTYON	2012	400000	573309	173309	43,3	0,72
CITROEN C4	2012	307700	437674	129974	42,2	0,71
RENAULT LOGAN	2012	236700	330565	93865	39,7	0,80

Заключение

В настоящей работе мы попытались разработать пробную версию алгоритма автоматического выявления наиболее выгодных автомобильных лотов на торгах по банкротству – инструмент, потенциально помогающий облегчить и ускорить процесс принятия решений для участников таких торгов. На текущем этапе алгоритм обладает некоторыми ограничениями: основные из них состоят в том, что признаки для базы 1, описывающие сами лоты, извлекались из текстов вручную, и в одной из построенных моделей эти признаки не дали приемлемого качества прогноза. Кроме того, до сих пор мы не учитывали в алгоритме ранжирования тех признаков, которые характеризуют сами торги. В связи с этим мы видим следующие шаги по усовершенствованию алгоритма и его дальнейшему масштабированию:

1. Создание модели для автоматического выделения признаков из текстовых описаний лотов на основе имеющейся размеченной базы 1.
2. Расширение базы с информацией о сроке продаже поддержанных автомобилей.
3. Создание алгоритма автоматического обновления базы 1, базы 2 и параметров моделей 1 и 2 (поскольку с течением времени цена автомобилей изменяется).
4. Добавление в процедуру ранжирования дополнительных признаков из базы 1 – например, характеризующих текущий этап аукциона, на котором находится лот, а также информации о состоянии лота и наложенных на него ограничениях.
5. Перенос всех этапов построения моделей на лоты иных категорий (прежде всего – недвижимости).

6. Оформление разработки в виде веб-приложения или бота, позволяющего реальным участникам торгов по банкротству отбирать наиболее выгодные лоты.

Источники

Chen, C., Hao, L., Xu, C. (2017). Comparative analysis of used car price evaluation models. *AIP Conference Proceedings 1839*, pp. 1-7.

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., Kevric, J. (2019) Car price prediction using machine learning techniques. *TEM Journal*, 8(1), pp. 113-118.

Google Trends [Электронный ресурс]. URL: https://trends.google.ru/trends/?geo=RU (дата обращения: 25.06.19)

Noor, K., Jan, S. (2017) Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*, 167 (9), pp. 27-31.

Oprea, C. (2010) Making the decision on buying second-hand car market using data mining techniques. *Special*, pp. 17-26.

Ozgur, C., Hughes, Z., Rogers, G., Parveen, S. (2016) Multiple Linear Regression Applications Automobile Pricing. *International Journal of Mathematics and Statistics Invention*, 4(5), pp. 13-20.

Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. *Advances in Information and Communication Networks*, pp. 413–422.

Pudaruth, S. (2014) Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, 4 (7), pp. 753-764.

Richardson, M. (2009) Determinants of used car resale value [Электронный ресурс]. URL: https://digitalcc.coloradocollege.edu/islandora/object/coccc%3A1346 (дата обращения: 25.06.19)

Shen, G., Wang, Y., Zhu, Q. (2011) A new model for residual value prediction of the used car based on BP neural network and nonlinear curve fit. *Proceedings - 3rd International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2011*, 2, art. no. 5721273, pp. 682-685.

Soejima, Y., Hirose, H. (2011) Auction Price Estimation for Used Cars by Regression Methods (Competition 1). *Proceedings of the Japan Society for Computer Science and Statistics*, pp. 9-12.

Sun, N., Bai, H., Geng, Y., Shi, H. (2017). Price evaluation model in second-hand car system based on BP neural network theory. *2017 18th IEEE/ACIS International Conference on*

Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 431-436.

Wu, J.-D., Hsu, C.-C., Chen, H.-C. (2009) An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications*, 36 (4), pp. 7809-7817.

Авто.ру [Электронный ресурс]. URL: https://auto.ru/moskva/ (дата обращения: 25.06.19)

Аукционы и торги по банкротству в 2019 году – самая подробная инструкция в интернете [Электронный ресурс]. URL: https://gdeikakzarabotat.ru/biznes/aukciony-i-torgi-po-bankrotstvu.html (дата обращения: 25.06.19)

Единый федеральный реестр сведений сведений о банкротстве [Электронный ресурс]. URL: https://bankrot.fedresurs.ru/TradeList.aspx (дата обращения: 25.06.19)

Мальцев А. Анализ торгов по банкротству и активности электронных торговых площадок (1 квартал 2018 г.). 2018. 33 с. [Электронный ресурс]. URL: http://download.fedresurs.ru/doc/BSR%20статистика%20ЕФРСБ%20с%20детальным%20рейтингом%201%20кв.%202011-2018.pdf (дата обращения: 25.06.19)

Федеральный закон от 26.10.2002 N 127-ФЗ (ред. от 29.05.2019) «О несостоятельности (банкротстве)» (с изм. и доп., вступ. в силу с 09.06.2019) // Собрание законодательства РФ. 2002. №43. Ст. 4190.

Электронные торги по банкротству (TBankrot) [Электронный ресурс]. URL: http://tbankrot.ru/ (дата обращения: 25.06.19)