

УДК 311.2

О НЕКОТОРЫХ ПОДХОДАХ В ОБРАБОТКЕ ПЕРВИЧНЫХ
ДАННЫХ ПРИ АНАЛИЗЕ РЫНКА

В. А. ГРИГОРЬЕВ, А. И. ПАНЬКОВА

Научный руководитель В. А. ЛИВИНСКАЯ, канд. физ.-мат. наук, доц.
Белорусско-Российский университет

Современный специалист для сохранения своего конкурентного преимущества должен владеть передовыми приемами сбора и анализа информации как в своей предметной области, так и в смежных областях. Особенно актуальными становятся эти навыки в эпоху возможности собирать данные в сети. Большинство из нас осуществляло такой анализ при выборе какого-либо товара, например, на сайтах интернет-магазинов. Потенциальный покупатель, в зависимости от критериев, которыми должен обладать товар, обязательно просмотрит несколько предложений и выберет оптимальный для себя товар. Это возможно сделать без особых затрат времени, если предложение не очень большого объема. В случае, если необходимо провести маркетинговое исследование рынка с целью моделирования цены предложения в зависимости от факторов, ее определяющих, то для получения более точного предсказания нужно учесть всю имеющуюся информацию.

Любое маркетинговое исследование начинается со сбора данных об интересующем объекте. Под данными понимается набор сведений, которые записаны на определенном носителе информации с целью их хранения, передачи для обработки и получения новых знаний.

Рассмотрим пример исследования формирования цены на вторичном рынке автомобилей Республики Беларусь в январе 2019 г. [1]. Цель исследования – получение индекса снижения первоначальной стоимости автомобилей различных классов, так называемого коэффициента амортизации, который широко используется в оценочной деятельности.

Ситуаций, когда необходимо использовать результаты такого анализа для предсказания рыночной цены автомобиля, бывшего в эксплуатации, довольно много. Например, при оценке ущерба в результате аварии для определения суммы выплат на восстановление авто. При различных юридических спорах, оценке остаточной стоимости основных фондов предприятия при банкротстве необходимо с наибольшей степенью достоверности получить эту величину. Для адекватного определения рыночной цены нужно собрать как можно больше информации о предложениях на вторичном рынке и проанализировать ее методами прикладного статистического анализа.

Такие исследования необходимо проводить регулярно, так как с течением времени рыночный спрос на подержанные автомобили

формируется под воздействием различных факторов. Главной проблемой получения коэффициента амортизации, основанном именно на анализе сложившейся конъюнктуры рынка, является сбор статистической информации. Современные информационные технологии позволяют автоматизировать этот процесс с помощью различных программ. Авторы исследования написали скрипты для парсинга данных с сайта ABW.by, являющегося площадкой для размещения объявлений продавцов и покупателей вторичного рынка автомобилей [1]. Было собрано и обработано 28 000 объявлений о продаже легковых автомобилей различных марок. Данные, собранные специально для проведения определенного анализа, называются первичными. Тип шкалы измерения позволяет классифицировать данные на количественные признаки (пробег, цена, объем двигателя) и качественные (тип двигателя и коробки передач, тип кузова). Для количественных признаков имеет смысл вычисление числовых характеристик, таких как среднее значение признака, дисперсия, стандартное отклонение. Для описания качественных признаков, которые в свою очередь подразделяются на номинальные и порядковые, определение среднего лишено смысла. Для номинальных данных, которые подразумевают принадлежность объекта к одной из нескольких категорий, может применяться только операция сравнения: «равно» или «неравно». Для характеристики совокупности по таким признакам применяется понятие «доля объектов совокупности, обладающих признаком».

В исследовании, проводимом авторами, согласно принятой классификации первичных данных, собранная информация представляет собой пространственно-многомерную выборку из векторов, элементами которых являются значения как количественных, так и номинальных признаков.

Так, факторами, влияющими на цену автомобиля на вторичном рынке, являются пробег, возраст, мощность двигателя. Эти факторы измеримы в числовой шкале, и автомобили можно однозначно ранжировать по ним и однозначно определить направление связи между ценой и факторами. Естественнo предположить, что с увеличением возраста автомобиля его цена на вторичном рынке будет снижаться. Методами эконометрического анализа можно количественно определить величину этого снижения с ростом возраста на единицу. А вот влияет ли на цену вторичного рынка цвет автомобиля, тип топлива, наличие дополнительных опций, тип кузова и как это влияние количественно измерить? При использовании такого типа данных в регрессионном анализе вводятся фиктивные переменные.

На первом, предварительном, этапе обработки первичных данных был произведен анализ исходной совокупности на предмет наблюдений, являющихся аномальными. Наличие аномальных наблюдений может сильно исказить результаты исследования, поэтому к каждому такому случаю необходимо подходить отдельно. В случае, если подтверждена

гипотеза о нормальном законе распределения исходной совокупности, для обнаружения единичных выбросов можно использовать статистические тесты (Z-value или Kurtosis measure, метод Тьюки, критерий Ирвина, критерий Граббса) [2].

В специализированных статистических пакетах реализована визуализация выбросов с помощью диаграммы «ящик с усами» с возможностью удаления их из дальнейшего анализа.

Если же при анализе аномальных наблюдений значительное количество, возможно совокупность не является однородной и необходимо применять алгоритмы кластерного анализа, который объединяет объекты в близкие по значению признаков в своей группе и максимально различающиеся с объектами в других группах.

В исследовании при анализе данных были обнаружены аномальные наблюдения по признаку «пробег» с помощью метода Тьюки. Поскольку объявления на сайт подают обычные люди, информация может быть искажена либо непреднамеренно, либо специально. Авторами был построен доверительный интервал для искусственно введенного показателя «отношение пробега к возрасту» и наблюдения, для которых этот показатель не попадал в него, были удалены.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Продажа и покупка автомобилей, автобизнес ABW.BY: объявления о покупке и продаже автомобилей. – Режим доступа: URL: <https://www.abw.by/>-. – Дата доступа: 20.01.2019.

2. Анализ данных: учебник для академического бакалавриата / В. С. Мхитарян [и др.] ; под ред. В. С. Мхитаряна. – Москва: Юрайт, 2016. – 490 с.