Особенности нейросетевого моделирования в задаче массовой оценки муниципальной недвижимости г. Москвы

К.К. Борусяк (*ФА при Правительстве РФ,* borusyak@gmail.com),
И.В. Мунерман (*ООО «Институт управления стоимостью»*, ivm@munerman.ru)

Основополагающую роль в управлении обширными комплексами недвижимости играет массовая оценка. Массовая оценка предполагает построение математической модели, позволяющей оценить с заранее известной точностью рыночную стоимость объектов определённой группы на основе ограниченного и фиксированного набора их основных ценообразующих характеристик.

В условиях кризиса рынка недвижимости данная задача приобретает особую актуальность при решении ряда проблем. Среди них мониторинг состояния залогов в банковской системе и их возможной эрозии (за счет обесценения), планирование инвестиционных программ государственных и частных институтов, деятельности строительных компаний и связанных с ними финансовых организаций. Преимущества массовой оценки перед индивидуальной во всех этих случаях достаточно велики: объективность, достигаемая за счет отсутствия человеческого фактора, возможность проведения оценки в режиме on-line с мгновенным пересчётом стоимости при появлении новых рыночных данных, относительно низкая стоимость внедрения по отношению к процедурам индивидуальной оценки.

Массовая оценка недвижимости является одной из наиболее классических для сферы экономики задач, в которых успешно применяются нейронные сети [3]. К другим из них можно отнести оценку кредитного риска и прогнозирование банкротства, стоимости производных финансовых инструментов, прогнозирование денежных потоков, в меньшей степени – прогнозирование доходностей финансовых инструментов и построение торговых стратегий (см. [5, 9]). Размерность факторного пространства высока, выборки достаточно объёмны, зависимости цен от факторов нелинейны и их форма неизвестна заранее – идеальные условия для нейросетевого моделирования. Обычные эконометрические модели (например, линейные или мультипликативные) в этом случае работают достаточно плохо: к примеру, в работе [4] аддитивная эконометрическая модель, полученная в результате редукции незначимых факторов, предполагает, что офисы (и объекты других назначений) класса B и класса D имеют одинаковую справедливую арендную плату.

1 Исходные данные и особенности факторного пространства

Хотя нейросетевые модели являются весьма эффективными в задачах оценки, их построение связано с двумя группами проблем, которые необходимо учитывать при предобработке данных. Во-первых, в отличие от ряда развитых стран (например, США, за исключением нескольких штатов, см. [7]), в России отсутствует система обязательного публичного раскрытия информации о сделках с недвижимостью, при которой сумма сделки и основные характеристики помещения, подлежащего продаже или сдаче в аренду, предоставляются в форме анкеты в соответствующие органы и агрегируются на открытых веб-сайтах. В связи с этим информация о сделках с недвижимостью крайне ограничена и не вполне достоверна - даже в Москве, а тем более в остальных городах России.

Для решения этой проблемы нами были применены несколько методов, что позволило существенно повысить качество исходных данных. *Семантические анализаторы*, основанные на регулярных выражениях, применялись для анализа текстов объявлений и выявления в них максимума информации, заданной в неформализованном текстовом виде. *Набор решающих правил* позволил исключить заведомо абсурдные анкеты, содержащие неправдоподобное сочетание признаков объекта недвижимости (например, офис класса А со стихийной парковкой). *Матрицы граничных значений*, составленные на основе эмпирических данных рынка недвижимости и статистического анализа выбросов, позволили отсечь объявления с заведомо недостоверной ценовой информацией.

Во-вторых, классические приёмы математического моделирования экономических процессов лучше всего работают в случае, когда все зависимые факторы являются количественными. В задаче определения цены объекта недвижимости факторное пространство устроено значительно сложнее. Большинство ценообразующих факторов являются неупорядоченными (например, назначение помещения) или упорядоченными категориями (состояние помещения – от аварийного до отличного). Важную роль играет также расположение объекта – географический фактор, кодирование которого представляет собой нетривиальную задачу. Простое использование географических координат не является решением проблемы, т.к. координаты – не ценообразующие факторы.

Первичный набор факторов, определявшийся экспертным путём с учётом наличия достаточного количества информации в основных риэлтерских базах, составил:

- выходная переменная: цена аренды или продажи объекта недвижимости;
- количественные факторы: общая площадь помещения (кв.м.);
- бинарные факторы: тип операции (аренда или продажа);

- неупорядоченные категории: назначение помещения (офисное, торговое, складское);
- упорядоченные категории: класс, состояние помещения, характеристика парковки и охраны, этажность (в виде текстового описания – «здание целиком», «со 2 по 4 этажи» и т.п.);
- географические факторы: расположение объекта.

Количественные факторы (с учётом преобразований, которые будут рассмотрены ниже) используются в модели в неизменном виде. Бинарные факторы задаются переменными-признаками: для типа операции «продажа» была выбрана за единицу, а «аренда» - за ноль. Неупорядоченные категории преобразуются в набор бинарных переменных, соответствующих всем уровням, кроме базового. В качестве базового назначения было выбрано складское (производственное).

Преимущество нейронных сетей перед моделями множественной регрессии состоит в том, что нет необходимости преобразовывать упорядоченные категории в набор бинарных переменных, теряя порядок значений, обусловленный экономическими причинами. Т.к. зависимости в нейронных сетях нелинейны, достаточно указать произвольные числовые значения, монотонно связанные с уровнями фактора, например, последовательные целочисленные значения или усреднённые значения цены в разрезе соответствующих категорий.

Расположение объекта было задано следующим набором потенциально ценообразующих факторов:

- престижность округа (неупорядоченная категория) – кодировалась при помощи рейтинга, рассчитанного исходя из средних значений цены в каждом округе при фиксированном назначении;
- расположение здания: на автомагистрали, оживлённой или удалённой улице (упорядоченная категория);
- расстояние до центра Москвы (измеряется от центра здания, в котором находится помещение);
- расстояние до ближайшей станции метро;
- расстояние до Третьего Транспортного Кольца;
- расстояние до ближайшей автомагистрали (крупного проспекта или кольца).

К некоторым из факторов были применены соответствующие функциональные преобразования. Цены и площади помещения были прологарифмированы. Кроме того, чтобы избежать разделения выборки на отдельные группы по аренде и по продаже, к арендным ставкам был применён коэффициент капитализации $1/r$, где $r = 12\%$ -

типичное значение валового рентного мультипликатора. Все факторы были нормированы путём вычитания минимального значения и деления на размах вариации.

Выборка составила суммарно 18 182 наблюдения. Она была случайно разделена на обучающую (80%), валидационную (10%) и тестовую (10%).

2 Архитектура нейронной сети: сети MLP и GRNN

В предыдущем разделе мы подробно рассмотрели исходные данные, которые легли в основу построения моделей массовой оценки нежилой недвижимости. Перейдём теперь к рассмотрению этих моделей. В силу описанных выше причин, методом моделирования стали искусственные нейронные сети. Они являются мощным инструментом для решения различных задач – распознавания, кластеризации, прогнозирования и др. Для каждой из этих задач существуют различные архитектуры нейронных сетей (виды нейронов, связей между ними, структуры сети), способы обучения сети и критерии оценки качества модели.

В задаче выявления зависимостей между переменными (регрессии) наиболее часто используются многослойные персептроны (multi-layer perceptron, MLP). Сети MLP состоят из входного слоя, на который подаются значения факторов, скрытого слоя и выходного слоя, на котором формируется результат. Настройка нейронной сети происходит путём оптимизации коэффициентов связи между нейронами с целью снижения средней относительной погрешности прогноза. Многослойные персептроны близки идеологии массовой оценки, т.к. позволяют выявлять глобальные закономерности в данных. Они являются нелинейными параметрическими моделями – обобщением регрессионных моделей.

Важной альтернативой сетям MLP являются обобщённо-регрессионные нейронные сети (GRNN, general regression neural network) [12], основанные на радиально-базисной функции (RBF). Такие сети успешно применялись в различных технических задачах, однако достаточно редко в сфере финансово-экономических исследований. К таковым можно отнести работу [9], в которой GRNN использовались для прогнозирования обменных курсов валют, а также работы [13, 14], посвящённые оцениванию систематического риска вложения в акции и оптимизации портфеля.

В задачах оценки рыночной стоимости недвижимости, насколько известно авторам, обобщённо-регрессионные сети ранее не применялись. В то же время, архитектура сетей GRNN (сильно отличающаяся от MLP) близка по идеологии к сравнительному подходу в индивидуальной оценке.

Сеть имеет один скрытый слой, и количество нейронов в нём совпадает с количеством наблюдений в обучающей выборке – сеть фактически запоминает выборку

внутри себя. Оценка стоимости рассчитывается как средневзвешенное значение выходного фактора (цены) по выборке, где веса определяются расстоянием между объектом оценки и нейроном. Чем ближе объект оценки к некоторому наблюдению из обучающей выборки, тем больший вес имеет соответствующий нейрон. Таким образом, обобщённо-регрессионные сети GRNN являются адаптивным и автоматизированным обобщением метода ближайших соседей, активно используемого в индивидуальной оценке. Сети позволяют оценивать стоимость объекта недвижимости на основе локальных особенностей факторного пространства, отдавая предпочтение близким аналогам, но используя информацию всей выборки.

Ещё одно важное преимущество сетей GRNN в управленческих задачах состоит в возможности определения и визуализации на карте объектов-аналогов, повлиявших сильнее всего на результат оценки. Это даёт возможность проверки адекватности модели (к примеру, можно убедиться, что расчёт для склада на окраине не основывается на информации о магазине в центре Москвы) и интерактивного повышения её точности. К примеру, если сеть MLP для некоторого объекта даёт неправдоподобный с экономической точки зрения результат, практически отсутствуют механизмы выявления причины и дополнительного улучшения модели. В сети же GRNN будет понятно, какой именно аналог (являющийся, скорее всего, выбросом) стал причиной ошибки, и его можно будет легко исключить. В то же время, к недостаткам сети GRNN можно отнести достаточно сильную чувствительность к выбору метрики исходных данных.

Хотя сети GRNN более релевантны управленческим задачам, априорно предпочесть GRNN и отказаться от MLP нецелесообразно. Каждая архитектура имеет свои преимущества и недостатки, поэтому выбор между моделями следует осуществлять на основе сравнения количественных критериев качества.

3 Качество моделей и полученные результаты

В качестве критерия качества моделей использовались среднеквадратические относительные ошибки (СКОО) прогноза на тестовом множестве. Т.к. выборка (даже после предобработки данных) потенциально содержит определённую долю выбросов, для вычисления СКОО необходимо использовать робастные оценки. В противном случае можно не только получить недостоверное представление о точности модели, но и снизить её, т.к. оптимизация параметров сети будет нацелена в большей мере на сглаживание выбросов, а не на повышение истинной точности прогноза.

В качестве оценки точности использовался нормированный межквартильный промежуток (см. [10]):

$$IQRS = \frac{\varepsilon_3 - \varepsilon_1}{1{,}349},$$

где ε_1 и ε_3 – нижний и верхний квартили относительных ошибок прогноза, а 1,349 – нормировочный коэффициент [8]. *IQRS* – один из робастных критериев, рекомендованных «Стандартом по анализу соотношения стоимостей» [1, пункт 5.4.2].

Сети MLP настраивались на обучающем множестве для различного количества нейронов, после чего оптимальный размер скрытого слоя определялся путём сравнения IQRS на валидационном множестве. Наилучшая модель MLP содержала 36 нейронов скрытого слоя, её средняя относительная ошибка составила 30,5%.

Аналогично, сеть GRNN запоминала обучающее множество, а затем проводилась оптимизация масштабирования факторов в целях снижения СКОО на валидационном множестве. Средняя относительная ошибка наилучшей модели GRNN составила 20,0%.

Для сравнения между архитектурами сетей (GRNN и MLP) использовался критерий IQRS для ошибок на тестовом множестве. Сравнение показало, что точность сети GRNN (как и на валидационном множестве) существенно выше: 20% против 35%.

Отметим, что линейная модель множественной регрессии позволяет достичь наименьшей погрешности 37,1% на обучающем множестве (на тестовом множестве ошибка возрастает ещё сильнее), хотя все коэффициенты значимы и их знаки соответствуют экономической интуиции. Более того, добавление в модель статистически значимых квадратов факторов и их попарных произведений позволяет снизить СКОО лишь незначительно (до 36,2%). При этом для обеих моделей (как до включения нелинейных по экзогенным переменным членов, так и после их добавления) RESET-тест Рамсея [11] отвергает гипотезу о верной спецификации модели. Это говорит о наличии существенной нелинейности в связях между эндогенной и экзогенными переменными.

Таким образом, регрессионные модели могут использоваться для верификации нейросетевых моделей, чтобы убедиться, что полученные взаимосвязи действительно имеются и носят нелинейный характер, а не являются результатом простой подгонки большого количества свободных коэффициентов синаптических связей нейронной сети. В то же время, высокая погрешность регрессионных моделей не позволяет применять их на практике.

Напротив, полученная погрешность в модели GRNN (20%) удовлетворяет Стандарту по автоматизированным оценочным моделям [1, пункт 8.4.5] и Стандарту по анализу соотношения стоимостей [2, табл. 1-3] Международной Ассоциации Налоговых Оценщиков.

Следует отметить, что разница в качестве MLP и GRNN может быть связана с тем, что помещения зачастую сдаются группами из похожих лотов – например, несколько офисов на разных этажах одного здания. Если один объект из такой группы попадёт в обучающую выборку, а другой – в валидационную или тестовую, локальное взвешивание аналогов в сети GRNN покажет мнимую высокую точность. В то же время глобальный характер обобщения в сети MLP может лучше описывать зоны факторного пространства, в которых наблюдений меньше. Возможно, было бы целесообразно комбинировать модели с различными архитектурами сети: при наличии достаточного количества близких аналогов использовать GRNN, а при их отсутствии – MLP.

По итогам создания модели для каждого назначения помещения были построены ценовые поверхности – карты, изображающие стоимость аренды или продажи стандартизованных помещений (например, класс D, площадь 150 кв.м. в хорошем состоянии на первом этаже здания, без охраняемой парковки) в зависимости от расположения в Москве.

Заключение

В данной работе рассмотрены подходы к осуществлению массовой оценки нежилой недвижимости - офисных, торговых и складских помещений. С учётом большого количества ценообразующих факторов, их сложной структуры, а также нелинейной зависимости между ценами и влияющими факторами, в качестве метода моделирования были выбраны нейронные сети. Настройка моделей на основе базы данных по сделкам с недвижимостью показала, что наилучшее качество показывает обобщённо-регрессионная нейронная сеть (GRNN). Этот результат согласуется с выводами работы [9], в которой проводится сравнение различных моделей для прогнозирования обменных курсов валют.

Среднеквадратическая относительная ошибка прогноза по модели составляет 20% - это типичная точность для моделей массовой оценки. Построенная модель позволяет повысить эффективность управления комплексами недвижимости в масштабах города или крупной корпорации и сделать этот механизм более прозрачным.

В то же время, существует ряд направлений совершенствования модели, прикладную ценность которых предстоит изучить в дальнейшем. Среди них можно выделить:

- включение в модель временного фактора для учёта и прогнозирования трендов на рынке недвижимости;
- точную географическую привязку объекта оценки путём включения в модель географических координат объекта в некоторой (например, полярной) системе;

- разработку механизма интерпретации результатов и определения основных аналогов, повлиявших на результат оценки, при использовании сети MLP;
- поиск оптимального комбинирования сетей MLP и GRNN в целях снижения общей погрешности;
- обобщение результатов на другие города России с учётом их особенностей и создание единой системы массовой оценки недвижимости в масштабах страны. При одновременном внедрении обязательного публичного раскрытия информации о сделках по аренде и продаже, это позволит перейти к налогу на недвижимость с его рыночной стоимости, о перспективах создания которого говорил Г.О. Греф [6].

Автоматизированная система массовой оценки на основе рассмотренной модели была разработана специалистами ООО «Институт управления стоимостью» и ЗАО «Производственно-коммерческая дирекция» и успешно внедрена в Департаменте имущества г. Москвы в 2008 году.

Литература

1. Стандарт по автоматизированным оценочным моделям. Международное общество налоговых оценщиков, 2003.
2. Standard on Ratio Studies. International Association of Appraisal Officers, 2007. http://www.iaao.org/uploads/StandardOnMassAppraisal.pdf
3. Аналитические технологии для прогнозирования и анализа данных. Учебник Нейропроект, 2005, http://www.neuroproject.ru/practice.htm.
4. Бывшев В.А., Богомолов А.И., Костюнин В.И. Оптимальное комбинирование прогнозов различных моделей массовой оценки стоимостных показателей объектов недвижимости. // Актуальные проблемы математического моделирования в финансово-экономической области: Сборник научных статей / Под ред. д.т.н., проф. В.А. Бывшева. Финакадемия, 2008. Вып. 7, сс. 23-37.
5. Бэстенс Д.-Э., ван ден Берг В.-М., Вуд Д. Нейронные сети и финансовые рынки: принятие решений в торговых операциях. – М.: ТВП, 1997.
6. Налог на дорогую недвижимость может быть введён после 2010 года. РИА Новости, http://rian.ru/politics/20070330/62824307.html.
7. Berrens R.B., McKee M. “What price nondisclosure? The effects of nondisclosure of real estate sales prices”. Social Science Quarterly, Volume 85, Number 2, June 2004, pp. 509-520.

8. Interquartile range. Help for Statistics Toolbox in MATLAB. http://www.mathworks.com/access/helpdesk/help/toolbox/stats/iqr.html.
9. Leung, M.T., Chen A., Daouk H. Forecasting exchange rates using general regression neural networks. Computers & Operations Research, 27 (2000), pp. 1093-1110.
10. Moore, D. S., McCabe, G. P. Introduction to the Practice of Statistics, 3rd ed. New York: W. H. Freeman, 1999.
11. Ramsey, J.B. Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 31, No. 2 (1969), pp. 350-371.
12. Specht, D.F. A general regression neural network. IEEE Transactions on Neural Networks, Volume 2, Issue 6, pp. 568-576.
13. Wittkemper, H., Steiner, M. Using neural networks to forecast the systematic risk of stocks. European Journal of Operational Research, 90 (1996), pp. 577-589.
14. Wittkemper, H., Steiner, M. Portfolio optimization with a neural network implementation of the coherent market hypothesis. European Journal of Operational Research, 100 (1997), pp. 27-40.