A Holistic Approach to Property Valuations

Using text analytics to detect crucial missing data from property listings can strengthen the property valuation process by ensuring greater accuracy.

Executive Summary

The unreliability of property valuations was one of the key findings that emerged from the (2008) Financial Crisis Inquiry Commission’s probe. Two key reasons among the various factors contributing to not-so-fair valuations are missing and/or inaccurate property listing data and not leveraging data from different sources to the fullest extent.

Incorporating all the available data (structured and unstructured) has its own challenges. With the current explosion of data, understanding the context and relevance and extracting meaning from the data available has become difficult.

An important step in the valuation process is identifying a set of comparables for a given “subject property.” Since the last recession, automated comparable identification methods have gained momentum given their consistency and lower costs. We believe that a holistic approach needs to be taken that optimizes various aspects such as selection of additional relevant data, and determining if this additional data can provide benefits and how to integrate the derived benefits into the business architecture. We use the sensing, thinking, acting and recursion (STAR) working model to approach the issue of ensuring more accurate valuations.

Introduction

Since the 2008 recession, property valuations have come under scrutiny in the mortgage business. In the light of unintentional as well intentional errors, regulators have recommended or mandated the use of more data from sources such as public records and property tax data in the valuation process. The challenge, however, is to store and analyze this additional data without compromising too much speed or accuracy. Also, it is important to understand the benefits that can be extracted from the additional data, since there is a belief that most of the data is redundant.

It has been estimated that only 20% of the data in the world is structured. A staggering 80% is unstructured data, most of which is not or cannot be leveraged for business purposes. The overwhelming rate of data generation is a challenge that most technology firms are trying to address. Another consideration is whether the bigger, unstructured data will actually lead to better results and will ROI actually increase proportionally with data.

The hype around big data seems to be subsiding and the focus has now shifted to trying to enhance the ROI of big data.
We looked at some real-world examples and case studies to understand the true benefits of using big data and allied concepts in the mortgage industry.

The case study below describes the use of text data to increase the accuracy of decision-making.

**Real Estate Industry Case Study: Using Unstructured Data for Enhancing Accuracy of Property Valuations**

Among the challenges faced by the real estate industry (especially for multiple listings) are inconsistency of data, incomplete/missing data and a multiplicity of sources. Multiple listing data is used to decide comparable properties (listed and sold) to value the “subject property.” Thus, data completeness is critical for identification of comparable properties and the valuation process. Comparable identification, in turn, is important for both manual appraisals (full or BPO) and automated valuations such as AVM. Since the 2008 recession, automated valuations and automated comparable identification methods/processes have gained some traction. These provide consistent results at lower costs, thus enabling more frequent valuations and better decision-making. The most frequently used automated comparable identification method is based on Euclidian distance on key property parameters such as area, age, number of bedrooms and bathrooms, etc. However, information on many of these parameters is missing in many multiple listings. The accuracy of valuations is not only relevant for identification of collateral fraud but also for providing an accurate picture to lenders on their individual property and portfolio losses. For example, a large lender in the U.S. has a portfolio of around 10 million properties. The average price of each property is about $250K. The typical default rate is around 10%. This works out to almost $250 billion worth of properties at risk. An error in valuation of only 10% would mean a potential loss of about $25 billion.

In such a scenario, leveraging unstructured data to improve the accuracy of valuation is well worth the effort.

We decided to analyze the data in eight multiple listings for various numeric, categorical and unstructured fields/variables. There were about 200 variables for which data can be captured by a real-estate broker. We noticed that the values in certain key variables such as garages were missing in up to 100% of some of the listings. We studied broker remarks (free-flow text data) to see if more information about these missing fields was available to validate the numeric data fields. We decided to use R code to find out in how many instances some information is provided in the broker remarks for the eight main variables: garage, appliances, living area, exterior features, construction material, parking carport, amenities and porch type.

**Provision of Information for Eight Main Variables**

<table>
<thead>
<tr>
<th>Field</th>
<th>Listing 1</th>
<th>Listing 2</th>
<th>Other 5 Listings</th>
<th>Listing 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garage</td>
<td>30% (100%)</td>
<td>4% (100%)</td>
<td>20% (81%)</td>
<td>71% (100%)</td>
</tr>
<tr>
<td>Appliances</td>
<td>10% (100%)</td>
<td>13% (38%)</td>
<td>15% (33%)</td>
<td></td>
</tr>
<tr>
<td>Living Area</td>
<td>10% (96.82%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exterior Features</td>
<td>0.20% (8.58%)</td>
<td>14% (46%)</td>
<td>2% (50%)</td>
<td></td>
</tr>
<tr>
<td>Construction Material</td>
<td></td>
<td>5% (15%)</td>
<td>5% (65%)</td>
<td>61% (98.76%)</td>
</tr>
<tr>
<td>Parking Carport</td>
<td>2.94% (98%)</td>
<td>5% (100%)</td>
<td>30% (100%)</td>
<td>49% (100%)</td>
</tr>
<tr>
<td>Amenities</td>
<td>10% (100%)</td>
<td>5% (100%)</td>
<td>13% (100%)</td>
<td></td>
</tr>
<tr>
<td>Porch Type</td>
<td></td>
<td></td>
<td>7% (81%)</td>
<td>49% (100%)</td>
</tr>
</tbody>
</table>

*Note: The figures in () indicate the missing % of the field in the entire listing. Those outside the parentheses indicate approximate % available from text remarks.*

Figure 1
Figure 1 (preceding page) indicates that text analyses can provide some information either for validation of the numeric/categorical variables or to fill in the missing data. This is different from the typical text analytics of sentiments and emotions in that it leverages much more granular information in the text. Such data/information in the unstructured fields has the potential to enhance the quality or quantity of the numeric and/or categorical field, thus enhancing the models and their lifts.

We conducted an analysis to understand the additional accuracy provided to the valuation models by virtue of additional data/info available through text data. We used a score developed to provide the closeness of the comparable properties to the subject property. We restricted the use of text analytics to broker remarks, though it can be used with any unstructured data.

Score Variability from Textual Analysis

The comparable property scores are a function of variables such as property type, area, number of baths, number of beds, price rate, proximity to subject, lot size, age, etc. A lot of information for these independent variables was available in the remarks field.

Figure 2 shows how the scores change when textual information is used to fill in missing data. There is a significant improvement in the scores for various properties - around 12% on average, and up to 30% in some cases.

Enhancing the Valuation Process

Since the last recession, several steps have been taken toward more accurate property appraisals. These include the setting up of the Financial Crisis Enquiry Commission, which identified property appraisals as a weak link in the mortgage process. This resulted in the formation of the Appraisal Commission.
Foundation authorized by Congress. It came up with several Valuation Advisories. A key aspect mentioned was that “using automation to select comparable properties that will produce credible and reliable value estimates is the challenge of the AVM.” On the other hand, it had been emphasized by Fannie Mae that existing automation software solutions are not sufficient and do not guarantee good data quality. As the identification of reliable comparable properties is important even in non-automated methods such as BPO/appraisals, any issue with data or the software will affect both automated and manual methods.

The leveraging of unstructured data becomes challenging particularly for mass appraisals. This is an area where the approach used in the case study above can be valuable. The entire process can be automated to a significant extent, thus meeting the regulatory requirements. In most real-world scenarios, there is significant redundancy in the data provided. This redundancy can be leveraged to validate and/or reduce missing values from data obtained from a single source or similar sources.

Below we discuss integrating these enhanced models with the existing selection process for comparables in terms of data, methodology and process.

Data: The STAR approach to addressing the valuation problem highlights that the agent remarks/comments in the multiple listing services (MLS) is one source of data regarding a property. The other possible sources are public records data, property tax data and property insurance data. However, data from each of these sources would have different limitations of accuracy. Demographic sources such as Acxiom, Experian, etc. can provide information about the neighborhood, etc. of the subject property. These aspects and challenges must be considered when designing the enterprise architecture.

Methodology: The Euclidean approach is a commonly used method to identify comparable properties in the industry because it is simple and intuitive. The overall methodology is not impacted by virtue of using the text analytics to validate/impute missing data. It is a step that comes before application of the selection methodology.

Process: The overall process of property valuation, either the application of automated valuation models to comparable properties or using comparable properties with respect to a subject for broker price opinion or full appraisal, requires comparable selection. The current process of comparable selection primarily uses MLS data. Software is used mostly to pull data from the listings and then to use it downstream, in the valuation process. The software deployed by the industry typically uses the RESO schema. One needs to integrate the text analytics approach used in the case study above with the pulling of the data from MLS and then move it to the data warehouse or update the database and have it reflected in the unified view of the data. In fact, developing and integrating a text analytics module to the existing MLS software provides a significant opportunity to the agent to be more effective in deciding which listing to focus on among the available set of listings. It also provides an efficient way for the end customer to extract more information from the listings quickly. Another challenge or opportunity is the development of software apps or platforms that are able to integrate data from non-MLS sources (property tax, public records, etc.) into the data warehouse.

The data model and the technology to pull data (structured, unstructured) from different sources would be context-specific. Based on our experience, however, we believe that a unified single view of the data is a significant improvement and aligns well with the modeling factory approach. This supports quick updates and validations of the valuation models.
Acknowledgements
Aan Chauhan, T.R. Rajarajan, Sunil Patil

Footnotes


14 Taggart, J., “Can you really trust just MLS data?” NDC Data Brief, 2013.

About the Authors

Edward Merchant is the Chief Technology Officer in Cognizant’s Banking & Financial Services Business Unit. He is responsible for advising and coaching BFS clients seeking effective and affordable ways to address chronic business and operational challenges through the creative use of both mature and emerging technologies. As the global co-lead for the BFS Technology and Architecture Office, Ed manages a team of solution architects and engineers responsible for converting concepts into implementable software designs. Over the course of his 30+ year career, he has held a variety of systems engineering, architectural design and IT operations leadership roles within financial institutions (regional and divisional CIO positions, Global Head of IT Strategy and Architecture, Global Head of Vendor Management), IT Services providers (sector and country BU Head positions) and strategic advisory firms (Big 4 partner). Ed holds an M.S. in mechanical engineering from Fairleigh Dickinson University and a B.S. in industrial education and technology from Montclair State University. He can be reached at Edward.Merchant@cognizant.com.

Nipun Kapur is Director and Chief Architect, TAO, in Cognizant’s Banking & Financial Services (BFS) Business Unit. He is responsible for specialized consulting to existing and potential BFS clients. Nipun has an M.Tech. and Ph.D. from Indian Institute of Technology, Roorkee. He has been associated with analytical/mathematical modeling for over 20 years - with the past 15 years focused on BFSI domains. Nipun has set up analytics COEs (GE Insurance Risk, CoreLogic) in India, leading highly qualified teams working on model development, evaluation, productization and production. He has been published in various national and international journals in the area of risk and modelling. Nipun is a certified Six Sigma (Green Belt) from GE. He can be reached at Nipun.Kapur@cognizant.com.

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world's leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 100 development and delivery centers worldwide and approximately 218,000 employees as of June 30, 2015, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.