ZILLOW - AN INDUSTRY PERSPECTIVE ON FORECASTING IN THE AGE OF BIG DATA

Dr. Krishna Rao, Director of Economic Research & Product @_krishnarao
The core of Zillow from the start was ‘Big Data’ and prediction

• The Zestimate - Zillow’s most recognizable product and the primary offering when the site was launched
  – A ‘now’ cast of what almost every home is worth produced using big data and techniques

• Layers of data and modeling to make the prediction
  – Data: Administrative county records + Multiple Listing Services + Census and other government data + User submitted data
  – Modeling: Tools and techniques from Econometrics, Statistics, Machine Learning etc.
Big Data at Zillow – Homes and Users

Living Database of 110M Homes
57M Home Profiles Updated by Community

- Buyers & Sellers
- Renters
- Homeowners
- Real Estate Agents
- Mortgage Providers
- Property Managers
- Home-Related Services

Consumer Empowerment
Through information transparency
1. Data
   - Sale Price
   - Last Sale Date
   - Square Footage

2. Info
   - Did housing prices go up last year?
   - Buyers’ or sellers’ market?

3. Insight
   - Zestimate
   - Home prices next year
   - Will that home sell above asking?
What makes ‘Big Data’ different for us?

• Partially it is about size (volume and velocity)
  – The methods we use need to scale
  – Real time computation in the cloud vs at the desk

• But many challenges come from how data is being generated (variety)
  – Then: Recording and collecting data was a very intentional task especially at large scale
  – Now: Increasingly data collection is the default - especially online

• Huge amounts of data
  – But that data can be less structured and translating to analysis can be more challenging
Examples of ‘Big Data’ we use

• Combine the original home attribute and price data with…

• User behavior on the site
  – Page Views, Saving a home, Clicking on a school boundary, etc.

• Images and videos of the property

• Unstructured listing descriptions of a property or reviews of an agent

• Etc.
<table>
<thead>
<tr>
<th>Year</th>
<th>Daily Valuations Created</th>
<th>Total Homes Covered</th>
<th>Unique models developed daily</th>
<th>Data consumed in valuation process</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>43 million</td>
<td>60 million</td>
<td>34 thousand</td>
<td>2 TB</td>
</tr>
<tr>
<td>2016</td>
<td>100 million</td>
<td>115 million</td>
<td>7.5 million</td>
<td>25 TB</td>
</tr>
</tbody>
</table>
An Application to Home Price Index (ZHVI) Forecasting at Zillow

- Produce thousands of monthly time series on home values at various levels of geography – from national to indexes for custom defined neighborhoods

- Also produce millions of forecasts monthly for each geography all the way down to the individual home level
  - Challenges around scale

Forecasting:

- Traditional approach: Pull standardized economic and other data at the relevant geography and fit a time series model

- How can we incorporate some of this newer ‘big data’ into our approach and modeling?
New types of data often require new methods

- Traditional statistical methods are very focused on the data generating process and how to make inferences around it
  - What is the target population?
  - What is the sampling process that generates the data?
  - What variables of interest do our measurements proxy for?
  - Etc.

- Much easier to think about those problems when your data comes from a well constructed research study than for data that comes from the online behavior of users who happen to visit a website for example

- That has partially driven the evolution of newer tools that are focused less on understanding the data generating process (or casual inference or model interpretability, etc.) but are more narrowly focused on predictive accuracy and have made amazing strides in accuracy on many tasks
  - Machine learning methods vs classic statistics/econometrics?

- IMHO: Forecasting and time series methods in particular have been a little slower to adapt new frameworks than most prediction problems in industry
Some ways we might use this data

- Utilize user patterns of search behavior to learn time varying interactions between regions
  - Is Harlem or Queens the new Brooklyn?
  - Fit a Gaussian Process to understand how the spatial correlations of search behavior and prices evolve over time

- Analyze pictures of homes to generate new potentially predictive features
  - Are Craftsman homes more in demand than Colonials?
  - Train a Neural Network to infer the style of a house and then add it as a regional fixed effect in our models

- Mine listing description to get a sense for the character of a neighborhood
  - Which areas show early signs of being in transition?
  - Identify areas in transition and correlate with terms from listing descriptions that might indicate hard to measure aspects of an area (family friendly vs tear down/bring your contractor)
Some examples of challenges we have faced

• Quickly run into the Fat Regression Problem: with all this data, many more predictors than observations
  – Methods to choose which variables to include (shrinkage, dimensionality reduction, etc.)
  – Need to careful about overfitting (cross validation and hold out sets)

• Hard to model in an environment that is always changing
  – How does a refresh of the site change behavior on it?
  – Training models on historical data is hard when the platform constantly evolves

• Integrating different modeling approaches is a challenge
  – Smart ensemble methods can pick from the best model in different areas

• Skill set required constantly evolving (and is in short supply)
  – Statistics, Economics, Machine Learning
  – Premium on statistical programming and data munging
  – Constant stream of new tools
Zillow is committed to freeing data whenever and wherever possible for two reasons:

1. We believe in **paying open data forward** and returning something to the open data community from which we benefit.

2. Free data is consistent with our mission to create a more transparent and well-functioning real estate market.
BIG DATA
And there are 2 million more...
CLARKE’S THIRD LAW:

Any sufficiently advanced technology is indistinguishable from magic.

[ Arthur C. Clarke ]