Executive Summary

HouseCanary is developing the most accurate and comprehensive rental valuations for residential real estate. Accurate valuations are the result of combining the best data with the best models. This whitepaper provides transparency to our performance and methodology used to achieve the most accurate nationwide single-family residential rental valuations.

The HouseCanary rental valuation algorithm is designed to use all available market data to estimate the most likely value that a property would rent for in the current rental climate for the market. HouseCanary strives to put a rental value on every single-family home, condominium, and townhouse in the United States. The value range should be as narrow as possible while still providing an approximate 68% coverage probability (one standard deviation) on actual rental homes listed in the MLS.

The purpose of this paper is threefold:

1. Identify the data included in the rental valuation algorithm
2. Provide context for the logic behind the rental valuation algorithm
3. Define key model outputs

At a high level, the rental valuation algorithm is comprised of two main steps: 1) First, we query and clean relevant data from HouseCanary's database, and then 2) train rental valuation models from historical rental prices. After the models are constructed, we apply these models to the broader population of homes.

HouseCanary rental valuation algorithm at a glance:

67M

U.S. residential real estate properties

All 50 states

Models for 3 property types

Updated monthly

7.1%

Median absolute prediction error (MdAPE)

Machine-learning-based

12 million non-owner-occupied homes valued

Macro- and micro-information considered

Table of Contents

Modeling Principles

Key to accurate valuations = comprehensive data + machine learning

Rental market analysis valuations often only consider recently listed nearby comparable properties in their determination of rental value. While recently-listed comparable properties are an important piece of the valuation puzzle, we can achieve better valuations by also considering all previous rental listings for the subject property and neighborhood properties, as well as information from multiple other sources: macroeconomic data, capital markets data, mortgage records, search and social data, and house/parcel data.

The primary assumptions behind HouseCanary's rental model are the following:

1. Rental list prices on the MLS are a good proxy for final contracted lease amounts. The HouseCanary model predicts what a rental property should "list" at under current market conditions.
2. Rental prices in a given neighborhood tend to move together through time.
3. There is a relationship between rental prices and home prices. The functional form of this relationship differs market to market and even neighborhood to neighborhood.
4. Machine-learning-based algorithms are better suited than classical comp-based models at recognizing and exploiting higher-order complex relationships among input variables and their relationship to the response variable.
5. Human effort should focus on enhancing existing datasets to generate better data and features, which can be fed into the algorithm, further improving rental valuation accuracy.

Data

Nationwide county assessor information, MLS, and other property-level data

Property-level data entering the rental valuation algorithm are composed of public record data, Multiple Listing Service (MLS) data, and other property-level information:

- **Public record data** includes property characteristics sourced from 3,100+ county assessor offices spanning __ million residential properties.
- **MLS data** includes property characteristics, rental-listing prices, and rental-leased prices where available. Often times, rental-leased prices are not reported back to the MLS. The most recent 3 years of rental-MLS listings are used for modeling purposes (Reference # of total listings considered here if it helps).
- **Other property information** includes data on property valuation, schools, market-level summary data, neighborhood prices, along with other details that impact value, such as proximity to a busy street/golf course or view quality.

Data from both MLS and public record are refreshed daily. The only potential delay in our dataset would be due to delays from the original source. As an example, if a particular MLS takes three months to update a rental listing to “leased/sold” status, we will not have access to the reported leased price of the home. Market-level summary data typically gets compiled and added into the database monthly or quarterly.

Ensuring Data Quality

Trusted data quality and certified by USPS CASS

HouseCanary achieves trusted data quality through systems and processes developed and supported by a team of data engineers, data analysts, and domain experts. Our system is designed for complete end-to-end visibility of data flows with layers of dynamic, intelligent controls.

Our process starts with full profiling of any data source that will feed into our platform. For the profiling, we use multiple full data instances over time to get a complete picture of content and expected changes. We create field-by-field, value-by-value rules that determine how the data maps into our content management system. Those rules, combined with intelligent monitoring, provide the first layer of control that flags suspicious data changes, anomalous and unexpected values, and inconsistent data use. Flagged content gets quarantined until approved as valid, the issue gets remedied by the source, or new handling logic gets implemented.

Once validated data flows into the content managements system, it gets linked and normalized at the address, building, and census levels. With the content linked, a second quality control pass uses multi-source comparison to arrive at a consensus view of the correct and usable data for a given object. Only this data gets fed into products and models.

We use a United States Postal Service Coding Accuracy Support System (CASS) certified service to validate, standardize, and match all addresses that feed into our system. This includes all addresses from data feeds and user inputs. All matching and standardization gets handled by this service. To protect against degradation related to updates, any change of any component of a full address triggers a new validation, standardization, and matching event for the subject address and all previously matched and related addresses.

Methodology

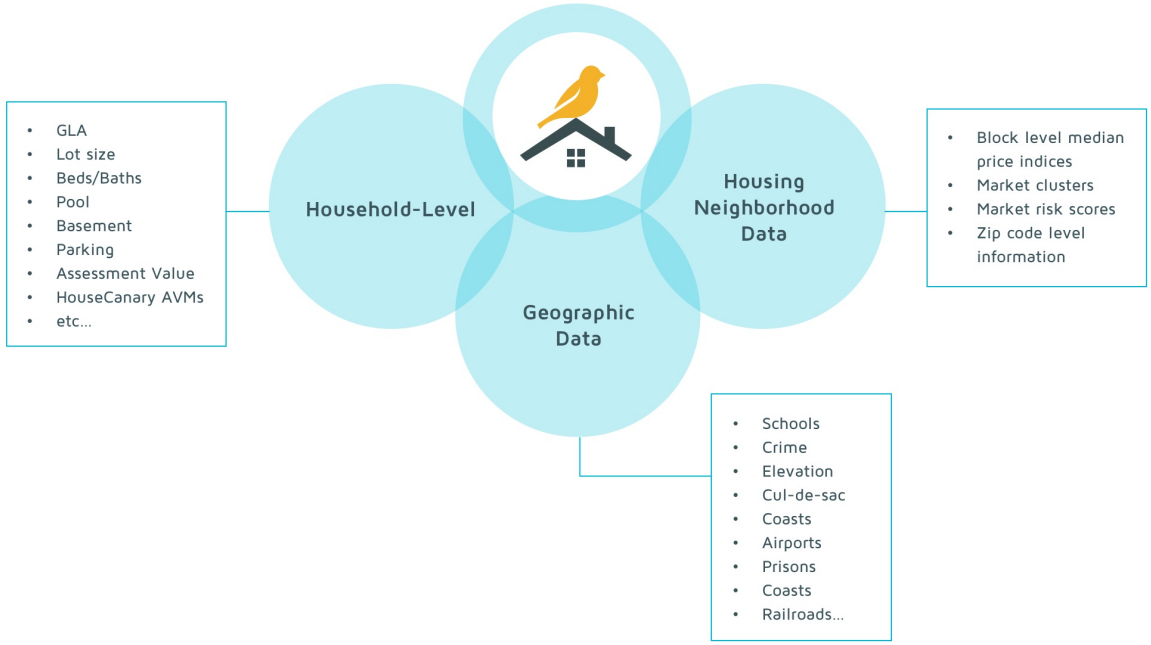
Two steps: 1) query and clean data, 2) train models

Step 1: Gather, combine, and filter data

Prior to modeling, the first step involves gathering, combining, and filtering all data we have access to. This step includes, but is not limited to, identifying valid rental listing from various MLS sources, identifying areas of limited data, cleaning various data sources, creating new features from the data, and determining valid property characteristics when data differs across sources. Listed rental prices and leased rental prices updated in the MLS form the basis for our response variable. Because it is often the case that rental listings are not updated with the contractual lease price, our final prediction aims to predict what a property should list for.

Because of the limited rental data in various MLS markets, models are trained at the national level. Many factors are used to capture the granular neighborhood effects. Running at the national level allows us to capture both macro- and micro-effects while machine learning allows us to account for unique neighborhood and market-level attributes. Exhibit 1 shows a small sample of some of the variables in which we include at the property, neighborhood, and MSA level.

Exhibit 1: Sample of property and geographic information considered in rental valuations



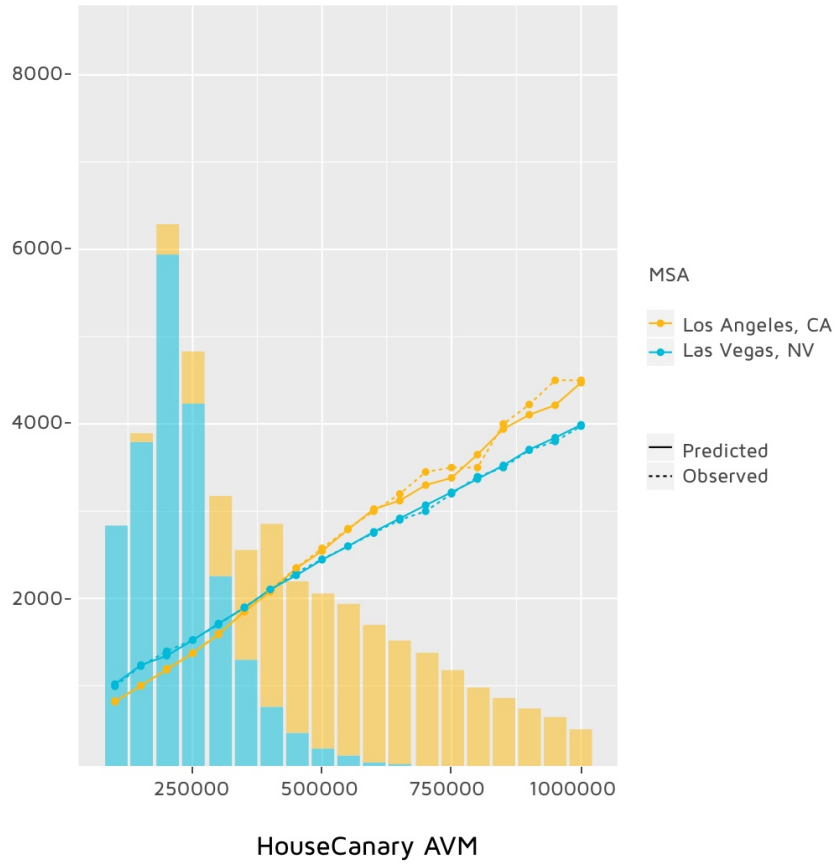
Step 2: Train rental valuation models

The second step involves fitting machine-learning-based models to explain the rental price from the cleaned data. In this step, we consider many different variables across many different data sources in order to determine a

rental list price. These available predictors include, but are not limited to: property type, fair market valuation (HouseCanary's AVM), block median price, time, property characteristics, neighborhood characteristics, macro- and microeconomic data, and spatial relationships.

Machine-learning-based methods allow us to account for both high-order interactions and nonlinear relationships between the predictor variables and the response variable. As an example, Exhibit 2 below shows the relationship of HouseCanary's AVM to rental list price in two different markets. Machine-learning models are able to capture the differences between these two markets in order to derive an accurate prediction of rental prices in both marketplaces. From the exhibit, you can see that Las Vegas' valuation distribution is very different from Los Angeles' price distribution. We are able to correctly capture the functional relationship between valuation and rental price in both of these markets.

Exhibit 2: Graph of rental price to building area for two very different markets: Los Angeles and Las Vegas



Machine-learning methods are also used to derive the upper and lower bounds around the rental AVM point estimate. Both the upper and lower bound estimates use all available variables that the point estimate had available in order to derive the range of uncertainty around the rental AVM estimate.

Machine-Learning Primer: Machine-Learning Applications for Real Estate Valuation

Real estate rental valuation involves predicting an output from one or more input(s), also known as a supervised-learning problem. The field of machine learning encompasses a set of algorithms designed specifically to tackle this type of prediction problem.

Machine-learning algorithms arrive at a solution by learning patterns from large amounts of data in order to make predictions. They seek to minimize the error from predictions by using information from many different inputs. The computer is presented with inputs and outputs. The goal is to learn a general rule, specified by the machine-learning algorithm, which maps the inputs to the outputs very accurately for the entire population of homes. The algorithmic process is iterative and often performs many iterations of error minimization in order to produce a robust and highly accurate prediction.

The patterns that exist in the data are not subject to the same parametric model restrictions as classic statistical models. The learning in the algorithm uncovers interactions among many different variables that cannot be functionally defined in traditional statistical models. The algorithms are naturally suited to capture both high-order relationships with the output variable and interactions among the inputs. Lastly, it is easily possible to consider many more inputs than actual observations. This means that a model may potentially contain thousands of inputs in order to achieve the most accurate prediction possible.

So what does all this mean for real estate rental valuation? If there truly exist predictive signals within the input dataset, these methods will find them with enough iterations. It is worth pointing out that the improved predictive performance of these methods comes at the expense of the additional computational time required to train them.

From these models, we generate an estimated rental price for every property in markets for which we have rental data coming from the MLS. Exhibit 3 shows how the final weighted historical rental price estimates compare to actual observed historical rental listings for the Los Angeles MSA. Exhibit 4 shows a histogram of the percentage differences between the actual listed amount and the predicted listing amount for the Los Angeles MSA as well. The error distribution, while not assumed to be normal, follows a roughly normal distribution here.

Exhibit 3: Actual vs. Observed Rent

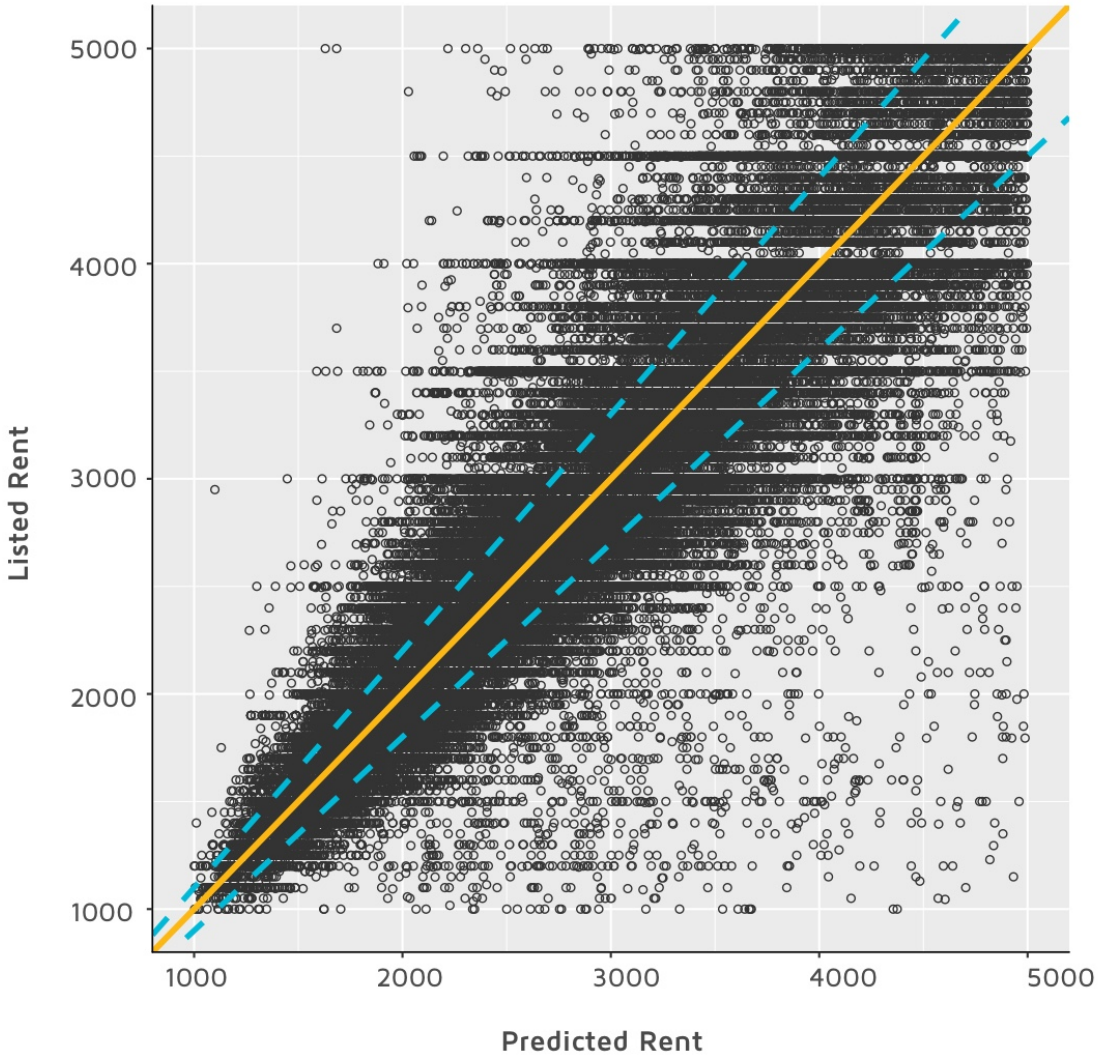
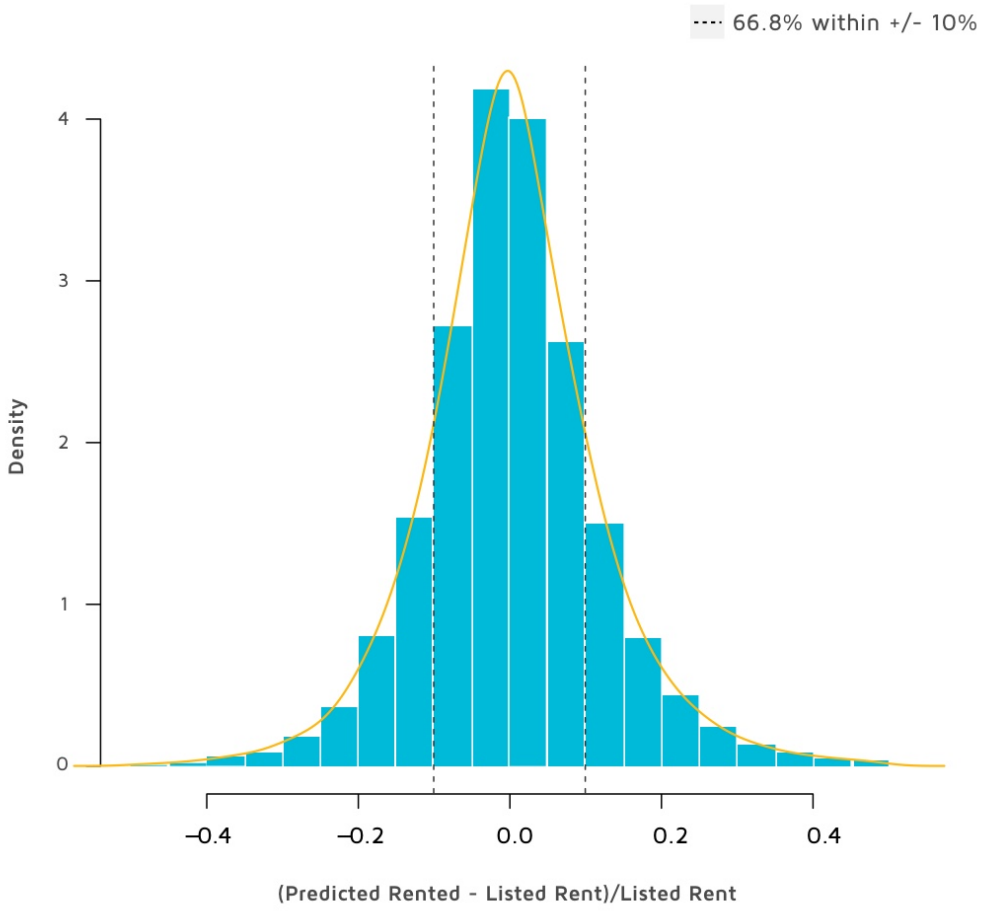


Exhibit 4: Histogram of percentage deviation from listed rental amount



Forecast Standard Deviation

Model-derived measure of uncertainty

The HouseCanary forecast standard deviation is a measure of model uncertainty. It is a quantity derived from the upper and lower bounds on the value estimate. The value range represents the quantity such that the range will actually capture the rental listing value in a rental listing approximately 68% of the time (one standard deviation).

The FSD is calculated as the percent of variation around the rental point estimate. It is calculated by (rental_upper – rental_lower)/(2*rental_avm). As an example, if the rental lower estimate is $800, the rental AVM is $1,000 and the rental upper estimate is $1,100, then the FSD is given by ($1,100-$800)/(2*$1,000)=0.15.

If a property were to be listed shortly after the estimate was generated, there is a 68% probability that the estimated range will cover the actual realized listed rental price. It is worth noting that the empirical error distribution is not necessarily normally distributed or symmetrical in our rental valuation model. Therefore, users

cannot obtain 95% coverage intervals by simply multiplying the FSD by 2 as with traditional statistical models. HouseCanary will make intervals of differing coverage probabilities available in a future release.

Another closely related quantity to the FSD is the confidence score. The confidence score is simply 1-fsd. If FSD equals 0.10 then the confidence score is 0.90. It is common to see both these quantities scaled up by a factor of 100, (i.e., 10 and 90).

Testing and Validation

Continuous internal testing

HouseCanary runs continuous testing on our rental valuation accuracy and coverage. All tests are composed of a six-month testing window. As an example, tests published on January 1, 2017 would be composed of a test sample over the period of July 1, 2016 to December 31, 2016. The six-month testing window is used to align with the testing window of HouseCanary's AVM.

Test accuracy attempts to measure how close a rental value estimate produced before an actual rental listing was to the subsequent rental listed price. As an example, let's say HouseCanary produced a set of nationwide property estimates using all known data through March 31, 2016. If the next scheduled model run takes place on April 30, 2016, then all rental listings occurring for the month of April would be compared to the most recently available rental value estimates prior to their listing date (the estimated values as of March 31, 2016). HouseCanary's ongoing internal tests track all of the performance measures below.

1. **'hit_rate'** – The proportion of listed properties in which we had an estimate of rental value prior to the listing.
2. **'Median_Abs_Pct_Err'** – The 50th percentile of absolute error in percentage terms. In other words, if this value equals 6.0%, then half our estimates are within +/-6.0% of actual listed price, and half are outside +/-6.0% of actual listed price.
3. **'Median_Pct_Err'** – The 50th percentile of actual error in percentage terms (not absolute error). Values close to zero imply that the estimator is unbiased.
4. **'Within X%'** – The percent of estimates that fell within +/-X% of actual listed price. HouseCanary produces this value for the 5%, 10%, and 20% bounds.
5. **'Within_HC_Prediction_Interval'** – The percent of actual listed prices that fell within HouseCanary's upper and lower estimates of value. The coverage probability of HouseCanary's upper and lower bounds is 68%. Therefore, this value should fall somewhere close to 68%, and values near 68% indicate that our model is accurately measuring the unexplained variance in price. As error rates continue to decrease, the width of the intervals will get smaller while still maintaining the target 68% coverage probability.

As of January 7, 2016, HouseCanary's continuous internal testing over the previous six months yielded a national MdAPE of 7.1% on 315,000 rental listings. Detailed test results, including the metrics above, are available by request at the national, state, and MSA levels. At a national level, HouseCanary's rental AVMs are also available at the property type and month level.

Future Implementations

Real-time valuation and updates to existing valuations

The current algorithm produces static nationwide estimates of rental value monthly. These are stored and used as our best estimates of value over the next month. Values are replaced once the algorithm runs again the following month and produces a new set of value estimates.

There are two areas for improvement with our current approach. First and foremost, new data received does not get considered until the next run. As an example, this could include additional MLS listings within the neighborhood, or updated property characteristics for one or more properties within the neighborhood which would yield to a more precise rental valuation. Second, we can only value properties that we have record of in our database. As an example, new homes may take 6+ months to be reported to us by the county assessor. In the current setup, we cannot value these until we learn of the property from the assessor 6+ months from now.

Future development will enable real-time rental valuations once a minimum set of information is entered about an unknown property. Minimally required input from the user includes an address that we can validate and the property type. Optional customer input will include corrected and/or previously unknown property characteristics. With the customer input in hand, a backend service will send the new information and generate an updated rental valuation.

About the HouseCanary Research Team

HouseCanary's research team is composed of PhD statisticians, economists, and mathematicians who use models and enduring data relationships to accurately see into the future of real estate. Using the latest in machine learning and predictive analytics, we build models that have worked accurately over the last 40 years.

Learn more about our research team and view their full bios at www.housecanary.com/about.

Contact

Please contact us with any questions or comments at sales@housecanary.com or **855.218.9597**.