Machine Learning Prediction Applied to Real Estate Investment

Summary

In this report, we apply machine learning to forecast Outcode Median prices one year in the future between 2005 and 2017. We find what is most significant is the robustness of the predicted rankings of Outcode performance (percentage price change over a one-year period) on a portfolio basis. Between 2006 and 2017[1] our Machine Learning predictions have a MAPE (Mean Absolute Percentage Error) of 9.90% with standard deviation of 1.75% versus Naïve predictions' MAPE of 10.80% with standard deviation of 2.22%. Portfolios based on the top ten Machine Learning predicted Outcodes return 13.83 times versus 0.74 times initial unit investment when compared to a portfolio of the top ten Naïve predicted Outcodes.

We then take predicted Outcode rankings and apply them to actual historical property price transactions as if they were traded through a property investment fund. We benchmark this return versus various in-class and other asset class returns. We show that our predictions and fund investment strategies outperform in-class and different asset class returns significantly, with a £50 million investment fund returning 3.74 times versus the House Price Index that returns 2.12 times. Furthermore, we show that on a simple strategy of buying in one year and selling the following year, the highest returns are associated with the Machine Learning predictions at a certain level of diversification when backtesting over the entire Land Registry repeated sales.

We also discuss how this stylised example can be adapted and operationalised in the real world.

[1] Excluding 2008

The hypothesis being tested is can one use machine learning models to predict real estate price movements and act on these signals in a way that is profitable. In this paper, we seek to show that it is indeed possible to accomplish this.

1. Data

We use data from the Land Registry, Ordnance Survey, Bank of England, Foreign Exchange and London Stock Exchange to create property price forecasts.

1.1 Price Paid Dataset

Of primary importance is the Land Registry Price Paid dataset. The Price Paid Dataset (Item 1) contains every property sold in England and Wales since 1995 and is updated monthly. The price paid dataset records the date of sale, the sale price, the type of property (Flat, Terraced, Semi, Detached), duration (Freehold, Leasehold), new build, address, postcode, unique transaction identifier and category record type (A/B).

1.2 Postcode Geocodes

Of secondary importance is postcode geocodes provided by the Ordnance Survey. These contain the Latitude and Longitude, or Northing and Easting of each full postcode in the UK. This data enables us to geocode the postcode location of properties for distance calculation purposes. Primarily this information comes from the Doogal London Postcode set (Item 2), which has the benefit of not only providing a full list of London postcodes, but also the updated geocodes from the Ordnance Survey. It is essential that any postcode geocode dataset contains both historical and current postcode geocodes because postcodes are continuously created and destroyed by the Royal Mail.

1.3 Other Data

We also use macroeconomic data, including interest rates (Bank of England Base Rate) and foreign exchange rates (Euro Sterling, Gold Sterling, Dollar Sterling) (Item 3) on a quarterly average basis. Lastly, we use FTSE 250 quarterly average prices (Item 4). These data sources provide a general indicator of the health of the broader economy (macroeconomic features).

2. Data Cleaning

The bulk of the data cleaning occurs on item 1. Here we work with the terraced property subset to deal with the problem of the heterogeneity of properties. Terraces have the second most abundant property transaction volume and are the most homogenous regarding the number of bedrooms and their sizes. Thus, when evaluating property prices, one can compare like for like properties which make for better forecasting predictions.

As well as selecting Terraces, we also filter for Record Type A only. Record Type A contains bona fide sales rather than Record Type B which include buy-to-lets, repossessions etc. (see Land Registry for more details (also further testing should include this different record type)). Finally, we do not distinguish between freehold and leasehold properties as we do not possess the lease length information to understand what discount a leasehold property would have over a freehold property (leasehold prices should be lower than freehold prices since the nature of a lease means the ownership time is limited, but in principle, unless the term of the lease is short <50 years, the price difference between a leasehold and freehold property is negligible).

All transactions without postcodes, or postcodes marked “UNKNOWN” are excluded as these will have no geolocation information and are negligible (<5) in number. Postcodes are then split into various categories to make the groupings of properties by their locations easier. The most critical postcode location resolution for

our purposes is the Outcode. The Outcode is the first part of the postcode and London contains over 250 Outcodes (current and historical).

Each transaction must then have a unique address id created to enable the vital step of matching repeated property transactions. Real Estate price modelling is fundamentally based on the principle of a repeat sales index, where repeated sales of a property provide the primary signal for how property prices are changing. Thus to create a repeats transaction database, we must be able to match properties that have more than one sale. To create the unique address id, every transactions PAON, SAON and POSTCODE are concatenated together. This is enough address data to create a unique address id for over 99% of properties that can then be matched later. Where this does not work is on the small number of properties where one address is many addresses, i.e. one house containing many flats, where each flat is a unique property that can be bought and sold but has no unique address details. This is more of a problem historically and in particular with flats, but these properties can be identified and removed during the cleaning process.

Item 2 is used to filter which postcodes are in London, and two different checks are made to ensure full London postcode inclusion. All postcodes that are marked as being in London in item 2 are included as well as those that also have the county as Greater London.

Moving on we proceed to remove duplicates and anomalous data from the records. First, the dataset is sorted by date such that we ensure duplicates are dropped in the correct time order (i.e. that the last record of any duplicate/semi-duplicate is kept). The first check for duplicated records is performed at the highest level, checking that no duplicates exist in Transaction Unique Identifiers (TUIs). Then one proceeds to check through every other category, holding all other categories constant. On cycling through New Builds, it is ambiguous whether a completely duplicated record was a New Build or not, so we decide to randomly select one of the records based on how they have been sorted. Next are Freeholds and Leaseholds. Here when we have a duplicate record, we favour the freehold transaction as it is more likely a terraced property would be a freehold than a leasehold (this would be opposite in the case of flats). This creates the bulk of the deduplication to ensure we have unique records.

Next, we proceed to create the repeats database that forms the basis of the backtesting strategy. Here we sort all transaction records by their unique address id, and then by date, as shown below:

Address 1 – Date 1
Address 1 – Date 2
Address 2 – Date 3
Address 3 – Date 4 etc.

This ensures that properties with more than one sale in the historical records are sorted in order of transaction. Records are then matched and shifted one period backwards, such that a paired list of transactions is created as below:

Address 1 – Date 1 : Address 1 – Date 2
Address 2 – Date 3 : Address 3 – Date 4 etc

All rows where address ids do not match are removed, and those leftover create the paired transaction database.

Finally, one further level of cleaning is performed, which is to remove quick/fast transactions, i.e. two sales within a period of three months. These may be considered as not being bona fide transactions, and so excluding them will help to have a more reliable understanding of price modelling. It may be said that some intermediate information is lost in this step, as shown in the following scenario:

Date 1 : Date 2
Date 2 : Date 3 (quick transaction)
Date 3 : Date 4

Would become:

Date 1 : Date 2
Date 3 : Date 4

Here intermediate information between Date 2 and Date 3 is lost, but what is important is that both Date 2 and Date 3 are still included in the set, and only the spurious information about the price change between them is excluded. It would be incorrect for instance to remove Date 3 and have a transaction directly between Date 2 : Date 4 as this would not be a realisable purchase and sale transaction when examining pairs in the pairs database.

The final list of property transactions is reassembled. This is formed of those properties that do not have any repeated sales, and transactions where repeated sales are not fast. This then forms the dataset to perform feature engineering on.

3. Feature Engineering

The model is set up to facilitate time series forecasting. As such a more conventional approach to Machine Learning modelling is taken whereby feature engineering takes place and then algorithms are deployed rather than performing deep learning on the cleaned unengineered/unstructured data.

The first issue to tackle is how to set up the problem to enable forecasting of prices sensibly. The target variable is defined as the median quarterly Outcode price one year in the future (or four quarters in the future). Property prices within an area can be quite heterogeneous, and there is considerable noise in the underlying signals. To smooth away the noise, a sufficiently large spatial resolution is taken in space and time, such that the underlying price movements, rather than the noise, are captured and can be forecast.

Each row in the dataset represents an Outcode and its quarterly median price. Because each row in the model does not have information about every other row, the relevant information for each row needs to be added to it. The most important features to add in for time series forecasting are the differences in price and the lags of the differences.

In designing the feature engineered dataset, some Outcodes are lost because they do not have a sufficient volume of transactions to offer a sensible underlying signal that can be predicted. This cutoff is defined as any Outcode that does not have more than 35 transactions in any quarter. Also, any transactions above £5 million are removed from the dataset as they are considered as outliers.

As well as price and differences in price features, various momentum, transaction, and area-based indicators are also incorporated. These include moving averages, nearest neighbour area prices, transaction volumes, number of new builds, Z-Scores etc. Item 2, item 3 and item 4 are added to the mix to provide each row with some contextual information as well.

Transformation of the time series was not explored at this stage. This would be included in further development and is discussed in the future AVM development plans section.

4. Machine Learning Approach

Training and testing sets are split into one-year forecast horizons between 2005 and 2016, with each subsequent year containing all previous training data. The price at the end of the following year is predicted as shown in Table 1:

Training Data Between	Test Data	Predicting Over (Year)
Q1 1995 - Q4 2004	Q4 2004	Q4 2004 – Q4 2005
Q1 1995 - Q4 2005	Q4 2005	Q4 2005 – Q4 2006

Q1 1995 - Q4 2015	Q4 2015	Q4 2015 – Q4 2016

Table 1: Machine Learning Training and Testing Setup

This then creates the forecast for the trading model to act on over the next year. At this stage, modelling in this way is sufficient for demonstration purposes of an investment strategy, and of course, better prediction will occur through modelling at higher time frequencies.

During Machine Learning we take a combined score from three different algorithms (Random Forest, Ridge Regression and XGB). The rank order of the Outcode predictions then forms the basis for input into the backtesting scaffold to calculate investment strategy returns.

Learning Parameters used are:

- Target Variable (Median Outcode Price one year in the future)
- Prediction Type - Regression
- Model optimized for RMSL (punishing predictions that are too large)
- Training Data split randomly with K-Fold testing (five folds – this has the robustness of meaning the model does not give greater weight to time periods closer in time to the prediction period and hence should be able to cope with unforeseen changes that occur in the prediction period)
- Use of XGBOOST, Random Forest and Ridge Regression Algorithms
- XGBOOST
    - Max no of trees - 1000
    - Max depth of trees – 15
    - Early Stopping
- Random Forest
    - Number of trees – 1000
    - Max depth of trees – 75
    - Min samples per leaf – 5
- Ridge Regression
    - Auto-Optimised

5. Machine Learning Results

The machine learning models generate RMSL errors of between 0.075-0.1 (MAPE of 4.5-7.5%) over the years tested. This then generalises to 6.4%-12.9% (financial crisis prediction year this becomes 30.2%).
Performance vs a Naïve Algorithm produces an error of between 8.4%-15.2% (also excluding 2008). This can be seen in Table 2.

Table 2 shows that the ML prediction errors are better than the Naïve algorithm in seven out of 11 years between 2006-2016. It is interesting to note that the Naïve Prediction errors are better than the Machine Learning algorithms during the years of and immediately after the financial crisis.

Years	ML Prediction Errors (MAPE)	Naïve Prediction Errors (MAPE)
2005	6.4%	
2006	**7.2%**	**10.1%**
2007	**8.8%**	**13.2%**
2008	30.2%	16.7%
2009	12.3%	9.3%
2010	10.3%	8.4%
2011	11.5%	8.4%
2012	**8.5%**	**9.0%**
2013	**9.1%**	**12.1%**
2014	**12.9%**	**15.2%**
2015	**9.9%**	**12.7%**
2016	**8.5%**	**9.6%**

Table 2: Machine Learning and Naive Forecast Errors

Each algorithm weights different features in the data:

- XGBOOST
    - Percentage Growth in London Price
- Random Forest
    - Moving Average Prices
- Ridge Regression
    - Closest Neighbours

Error distributions during training appear to be quite normal.

The reason these algorithms were chosen was due to their predictive power and the fact they attach importance to different parts of the data provided. One may ask why a deep learning approach is not taken at this stage, and this was because of the amount of data available (not enough) and the time required to train deep nets. Initial neural net testing did not show any improvements over the above algorithms. With more data and greater resources available, deep learning may show better predictive power, especially with the inclusion of further property level and contextual data sources. More about this can be found in the future Machine Learning plans section.

5.1 Machine Learning Results – Outcode Returns

To see how effective these predictions would be in the creation of a property portfolio, one would need to buy properties within the Outcodes. Thus, the Median Outcode Prices that we model serve as the unit we are buying in that Outcode (i.e. the median price property) and the return on that property would serve as the percentage price change annually.

In the creation of a portfolio, if we invest an equal unit of risk in each Outcode, then the return of a portfolio is simply the average of the returns from each Outcode. We test several strategies in evaluating portfolio returns by buying the top ten Outcodes ordered by their forecast growth. The returns on the Machine Learning, Naïve, Anti (the reverse ordering of the ML predictions) and Market returns (Market returns are the average return from all Outcodes) are shown in Table 3.

Year	% Returns on Investing in the Top 10 Outcodes (Except Market)				Compound Returns of 1 unit Invested into the Top 10 Outcodes (Except Market)			
	ML	**Naïve**	**Anti**	**Market**	**ML**	**Naïve**	**Anti**	**Market**
2005	11%		-4%	2%	1.11	1.00	0.96	1.02
2006	32%	9%	4%	11%	1.47	1.09	1.00	1.13
2007	25%	12%	15%	16%	1.83	1.22	1.15	1.31
2008	5%	-19%	-19%	-12%	1.92	0.99	0.93	1.15
2009	38%	-10%	-14%	6%	2.65	0.89	0.81	1.22
2010	30%	0%	-4%	8%	3.44	0.89	0.78	1.32
2011	19%	-15%	-13%	1%	4.10	0.76	0.68	1.33
2012	35%	-3%	-10%	8%	5.52	0.74	0.61	1.44
2013	35%	6%	0%	13%	7.42	0.79	0.61	1.62
2014	30%	5%	1%	18%	9.67	0.82	0.62	1.91
2015	24%	-2%	-4%	12%	11.96	0.81	0.59	2.15
2016	16%	-8%	-8%	6%	13.83	0.74	0.54	2.28

Table 3: Percentage Returns and Compound Returns of investing in the Top 10 Outcodes by predicted growth

From Table 3 we see that the returns in the Top 10 Outcodes are best in the Machine Learning portfolio, followed by the Naïve portfolio and finally by the Anti portfolio. On a compound return basis, the ML portfolio outperforms the market by approximately six times and both the Naïve and Anti portfolio return less than their original investments. These returns can also be seen in Figure 1.

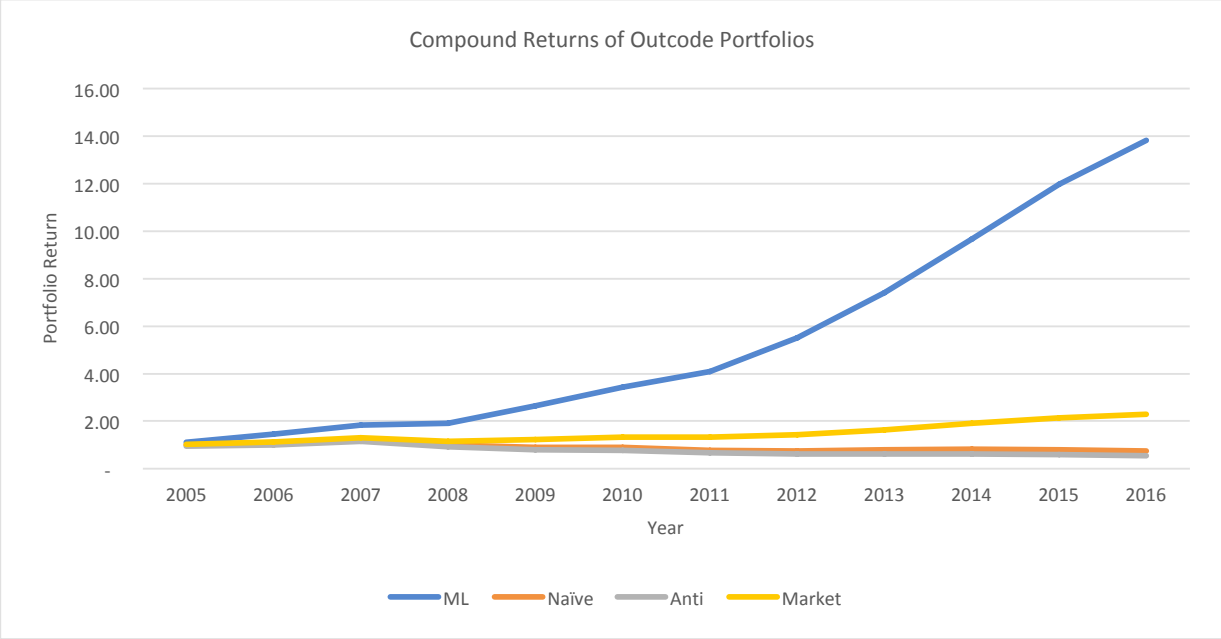


Figure 1: Compound returns of the Top 10 Outcode portfolios by prediction strategy

It is interesting to note that although the prediction errors may have been better on the Naïve portfolio in the years during and immediately after the financial crisis, the returns of the Machine Learning (ML) portfolio on the top ten Outcode portfolio are much greater. Here we see that more interesting than the Outcode predictions themselves are the orderings of the Outcode returns. The ordering of returns in the Outcodes shows much greater stability in their returns versus the Naïve and Anti portfolios that quite quickly

tend to the market return. This can be seen for 2006 in Figure 2. Here the Market return is the ordering of Outcodes by the return from highest to lowest.

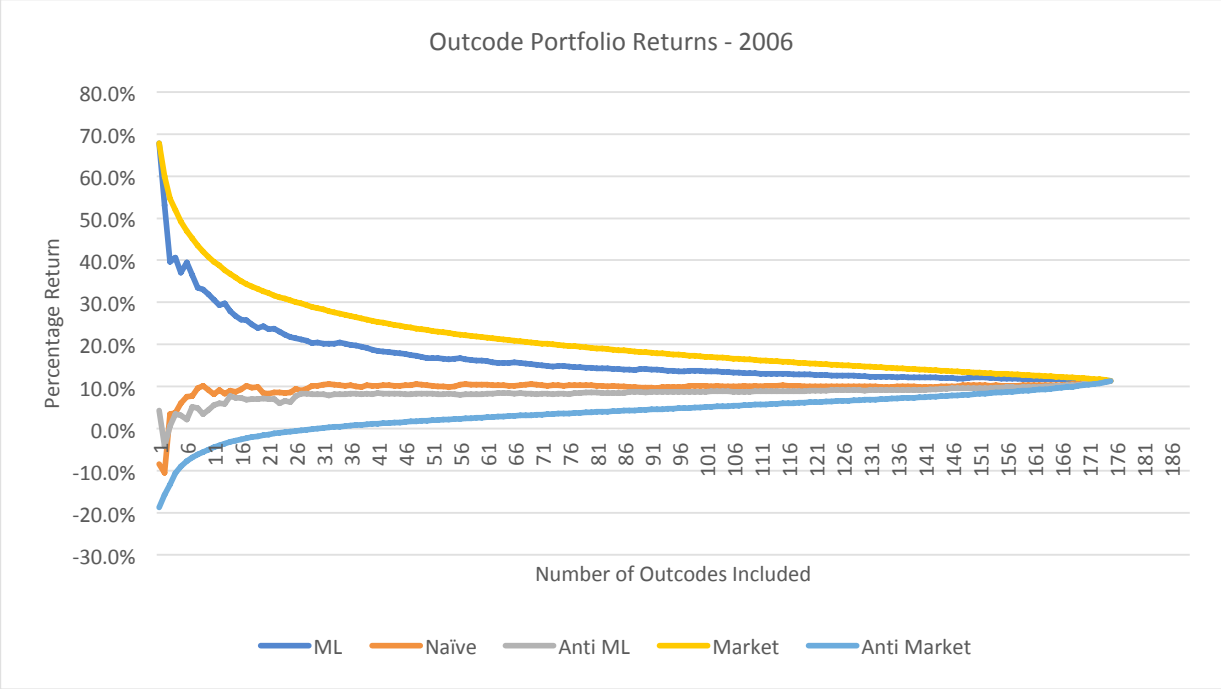


Figure 2: Average returns by number of Outcodes included in a portfolio (Number of Outcodes versus Percentage Returns)

In each year, the ML returns mirror the orderings of the Market with some examples of noise within the top five Outcodes selected but tending slowly to the market return from a position of outperformance versus the Naïve and Anti portfolios that tend quickly to the market return from a position of underperformance. Other years' results can be found in Appendix 1.

6. The Repeats Sample From The Land Registry Data

To backtest an appropriate buying and selling strategy, one must look to bona fide repeat sales within the Land Registry data. This means that we must restrict properties looked at to those that have at least two sales in the historical records. This can then be used to calculate profits or losses on real sales at verified prices. The alternative is to use an AVM to calculate backtested selling prices on any property that was sold. This adds an element of error into the backtest because you are relying on an AVM calculation for the selling price that contains a margin of error (and are an estimate). If a conservative AVM calculation is used to revalue properties, then this may be acceptable.

Instead of introducing any margins of error, we rely on the Repeats Database as we believe that there can be no argument about the transaction prices that took place in the market if this dataset is used.

It is important to understand how similar the Repeats sample is to the entire Land Registry Dataset given it is a sample of the Land Registry. We can look at several different statistics to help us determine this.

The first is Price Volume. This is simply the sum of all transaction prices. As can be seen in Figure 3, the Repeats sample is broadly aligned with the Land Registry set. The Correlation between the Land Registry and the Repeats Sample is 0.97. Also note the behaviour of the Repeats Buy Price Volume and Repeats Sell Price Volume coupling and decoupling from the repeats sample as we progress through time.

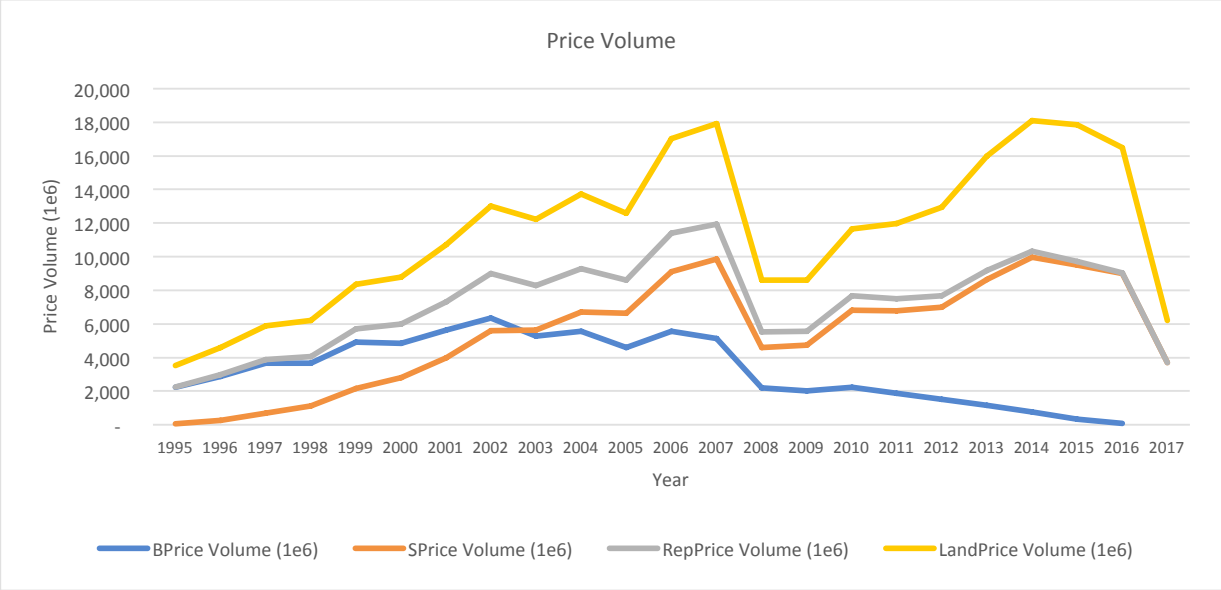


Figure 3: Price Volume comparison between the Land Registry set, Repeats Sample, Primary Repeat Transactions (i.e. the buying price) and Secondary Repeat Transactions (i.e. the selling price)

Next, we look at the number of transactions between the Land Registry and the Repeats sample in Figure 4. Both broadly follow each other, and the correlation is above 0.98. As expected the number of Buy Repeat Transactions tends to zero as we approach current time, whilst the number of Sell Transactions becomes the only constituent of the Repeat Transactions.

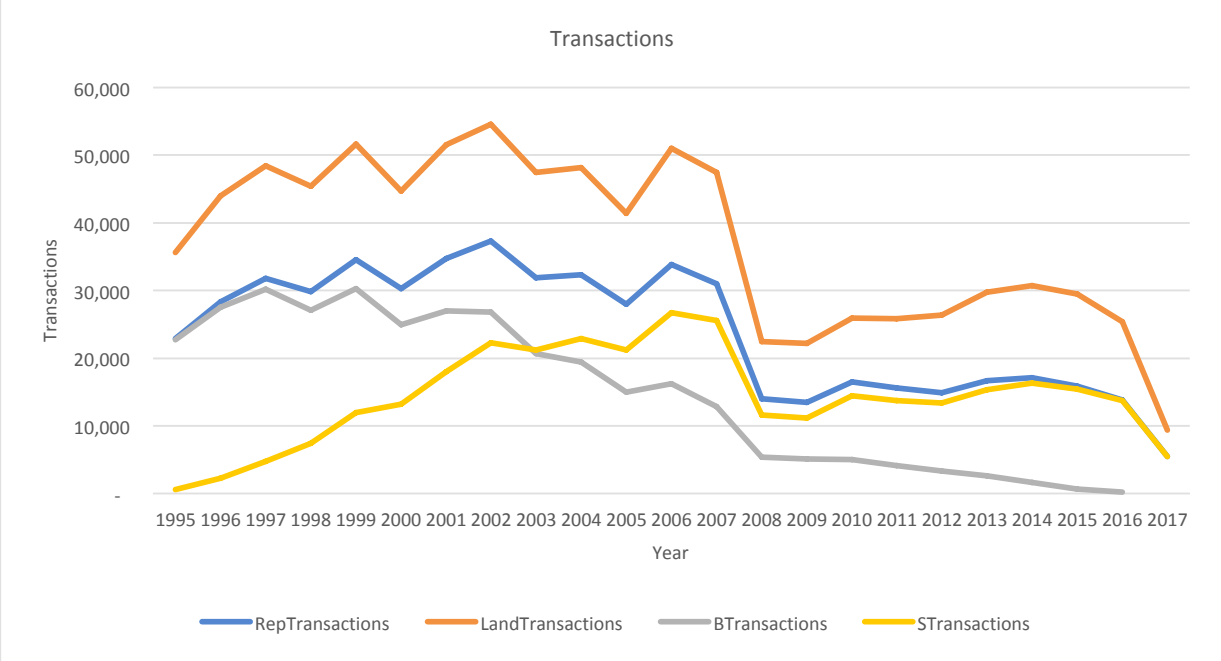


Figure 4: Number of transactions in the Land Registry, Repeats Sample, Repeats Buy Transactions and Repeats Sell Transactions

Figure 5 shows the Median Prices between the datasets. The Land Registry and the Repeats sample have a correlation of 1. In the case of the Buy and Sell Median Prices we see the Selling Median Price is lower than the Buy Price and that the Buy Price becomes lower than the Land Registry and Repeats Sample after 2012. The Buy Price decouples from the Land Registry and Repeats sets after 2012 because we approach the

threshold of the average hold duration of a repeat transaction which is approximately 5.5 years. After this point, you are primarily dealing with fast transactions, usually where value is picked up in flipping (rapidly buying and selling) a property. These properties tend to be below the median price.

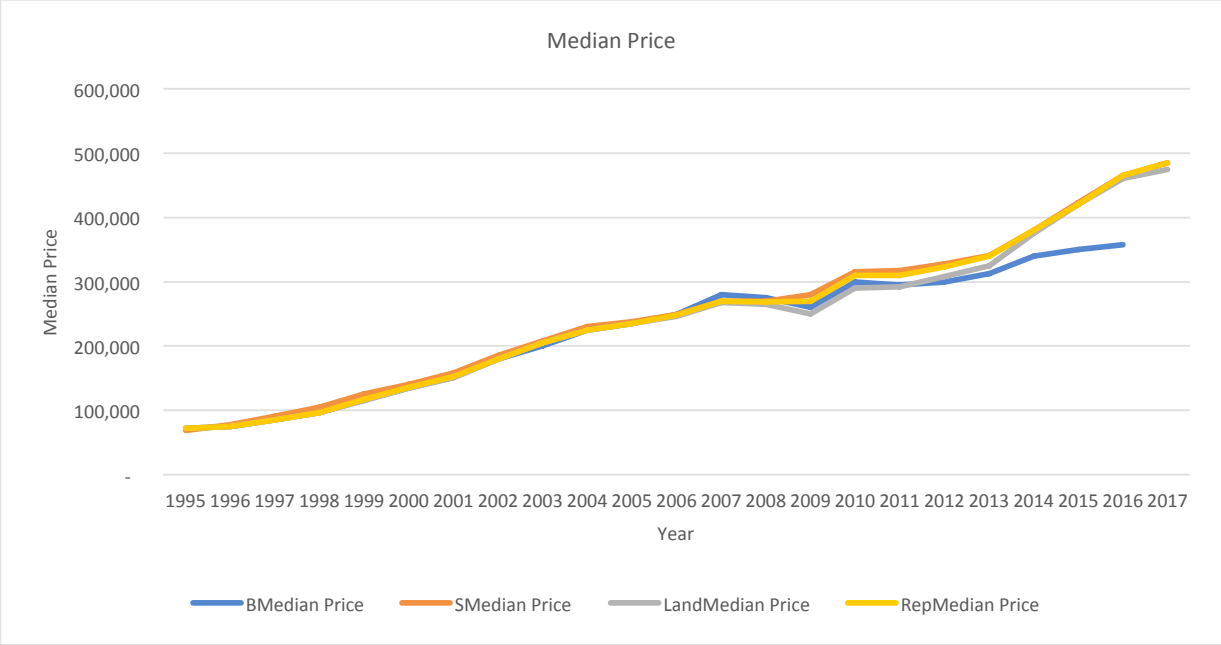


Figure 5: Median Prices of the Land Registry versus the Repeats Sample

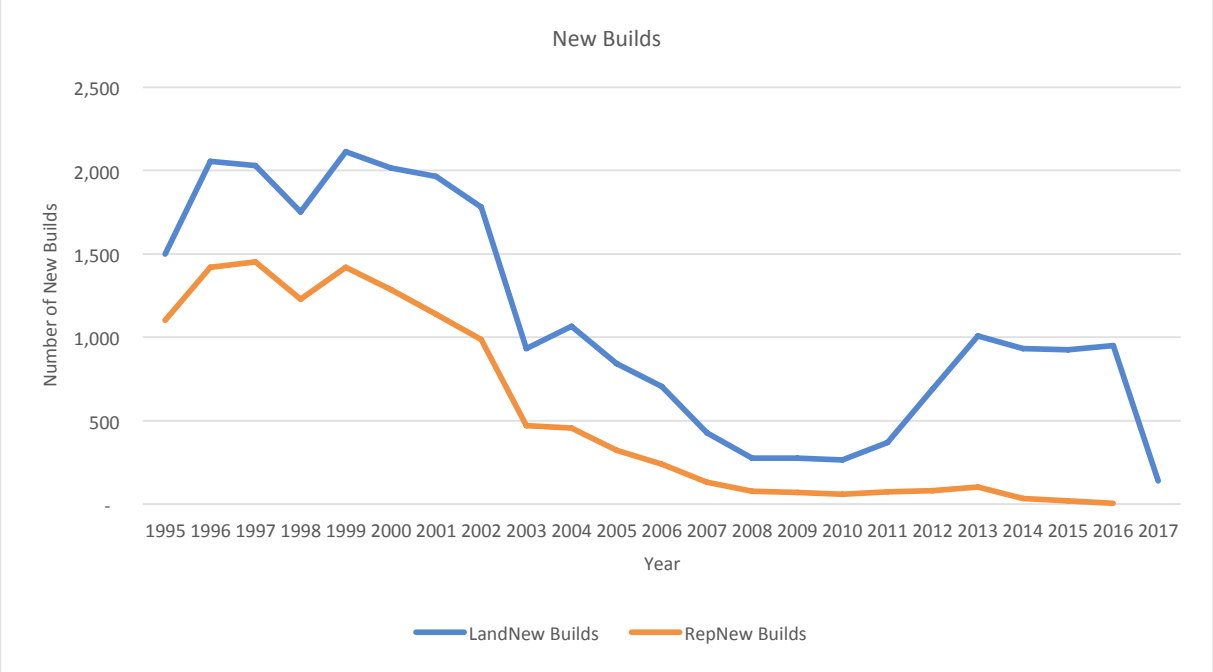


Figure 6: Number of New Builds

Figure 6 shows how the number of new builds in each of the samples changes. The correlation between the Land Registry set and the Repeats set is 0.93. There is some decoupling after 2011/2012, but once again this is due to the effect of the average hold duration of repeat sales being seen at the end of the sample. Here we see that new builds that have transacted have not come back onto the market yet.

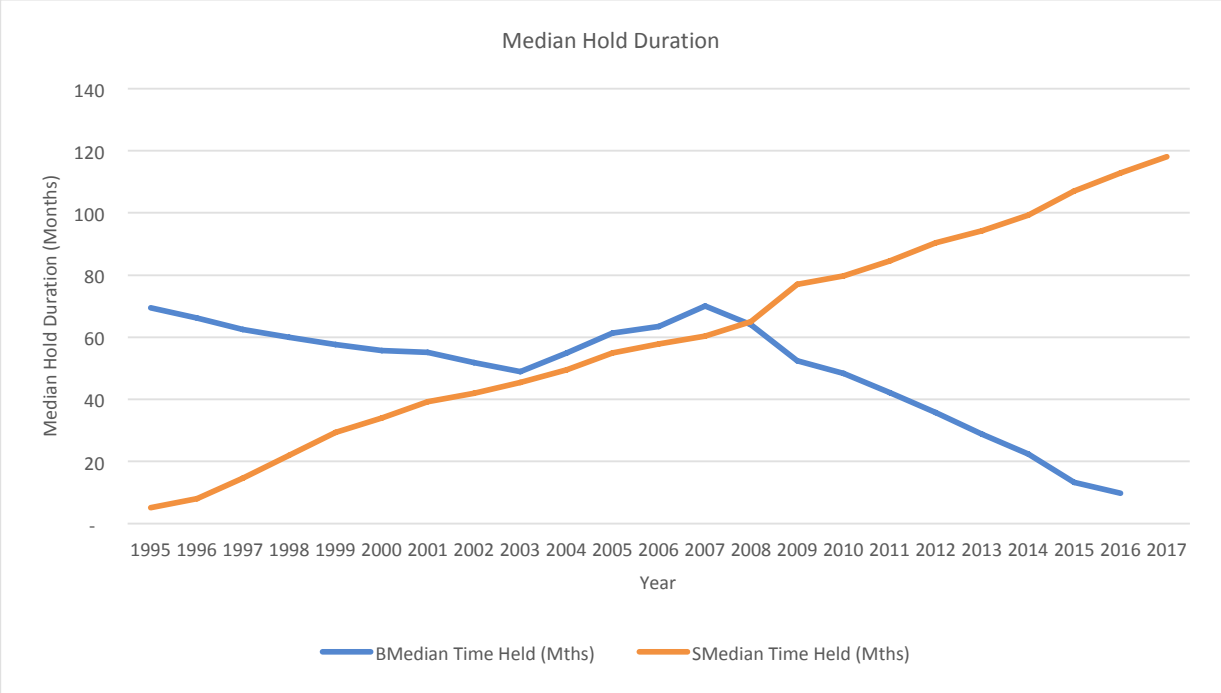


Figure 7: Median Duration for Buy and Sell Repeat Transactions

Figure 7 shows the median hold duration for a repeat transaction. For the entire repeats set, the median hold period is approximately 5.5 years. In the above figure we see that selling and buying hold times are negatively correlated, and as expected, buying hold times tend to zero as we reach current time whilst selling hold durations continue to increase.

It is also worth exploring the distribution of prices between the Land Registry, Repeats sample and Buy Repeat and Sell Repeats set. In Appendix 2 the distribution is cut off at £1m, and you can see that there is a long positive tail. All the samples mirror the Land Registry broadly (mean, median, standard deviation), and you see the Sell Repeats become the Repeats set as we reach current time and the Buy Repeats volume tends to zero. As time progresses, we see the mean and median prices shift rightward.

From the examination of the Land Registry and Repeats sample we can state that testing on the repeats sample will hold to the general Land Registry set. The only point of concern would be transaction volumes tending to zero as we approach current time. This is discussed further in the next section.

7. Backtesting

Now that the use of a repeats sample has been discussed for backtesting, we can describe the methodology used to transition from the Median Outcode Price Returns testing to actual testing on property transactions.

Running through strategy simulations (on 100 simulations to account for the nature of the random purchase decisions made in the Outcodes (by way of the random seed)), the fund return and asset returns on a £50m portfolio are presented in Figure 9.

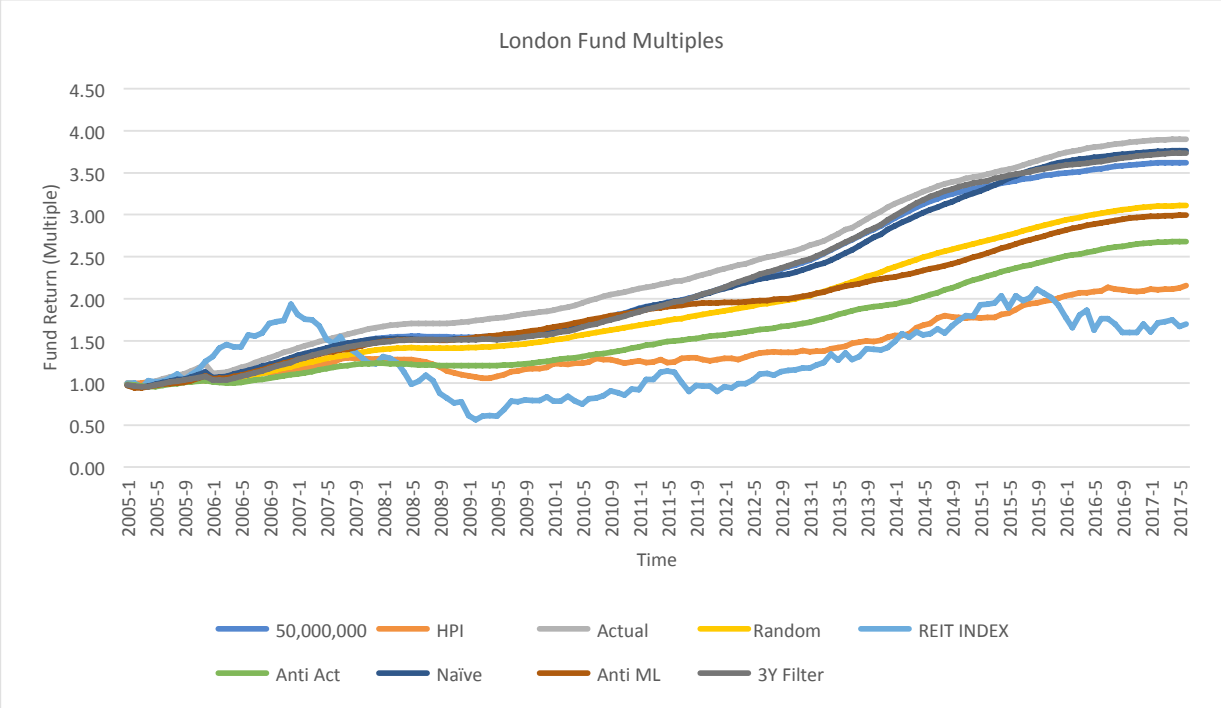


Figure 8: London Fund Multiples (All)

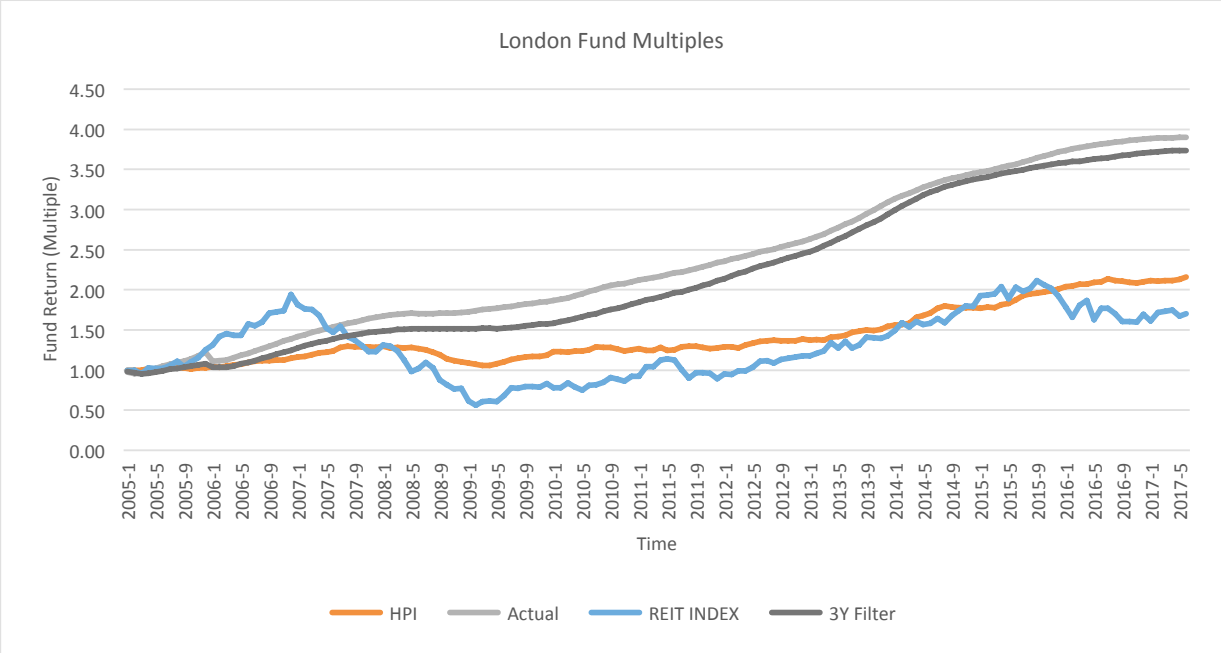


Figure 9: London Fund Multiples (Selected)

We can see that the ordering of the strategies intuitively matches (except for one) the ordering of the fund multiples, i.e. knowing the actual outcomes produces the best results, whilst the naïve strategy and our

Machine Learning (3Y Filter)[2] strategies come second and third, followed by the random strategy and finally the worst is the anti-strategy (here the anti-strategy is defined as the reverse order of the actual results).

We see that the Naïve and Machine Learning strategy closely track each other with the Machine Learning (filtered) strategy, in fact, outperforming until the second half of 2015 where the Naïve strategy overtakes this. This is further discussed in a later section of this document, but for now, the outperformance of the Naïve may occur due to the choice of variables and the fact that the Machine Learning predictions are only calculated on an annual basis whereas other strategies update on a quarterly basis.

It is interesting to see as well that even the worst strategy, the anti-strategy can produce a return that is greater than both the REIT index and the HPI (where the HPI does not account for fees). One may think this is a consequence of the repeats transaction database, but as we have shown above the repeats sample is well aligned with the Land Registry set. This most likely occurs because we are using a trading strategy, whereas the HPI is a buy and hold strategy.

The fund returns themselves are affected by items such as the level of cash at any point in the portfolio that is not being invested and hence earning a return. To evaluate the true returns on assets, one can instead look to the absolute asset return as a means of assessing the effectiveness of a strategy.

Figure 11 shows the assets returns, and you can see the differences between the strategies are subtly defined showing what you would expect regarding the ordering of the strategies. The differences in the asset returns are not as clear-cut as with the fund returns after approximately 2009, but the nature of supply and demand, as well as asset returns, may be accounted for in the strategies which are why you see a more explicit outperformance when looking at funds (i.e. a lower asset return with more transactions can outperform a higher asset return with fewer transactions on aggregate). It is also interesting to see that the Naïve portfolio outperformance of the filtered Machine Learning strategy can be seen through Asset returns in the period after the second half of 2014, but that its quite large underperformance is witnessed during the financial crisis. This is clearly a sign that a pure momentum play enables you to capture a great deal of value in the repeats sample, even though the Naïve model does not perform under the Outcode Median Price testing environment.

[2] The Machine Learning Filtered (3Y Filter) strategy applies a confidence feature with a short-term memory, where predictions are reordered based on the accuracy of recent predictions (further discussed in future AVM development)

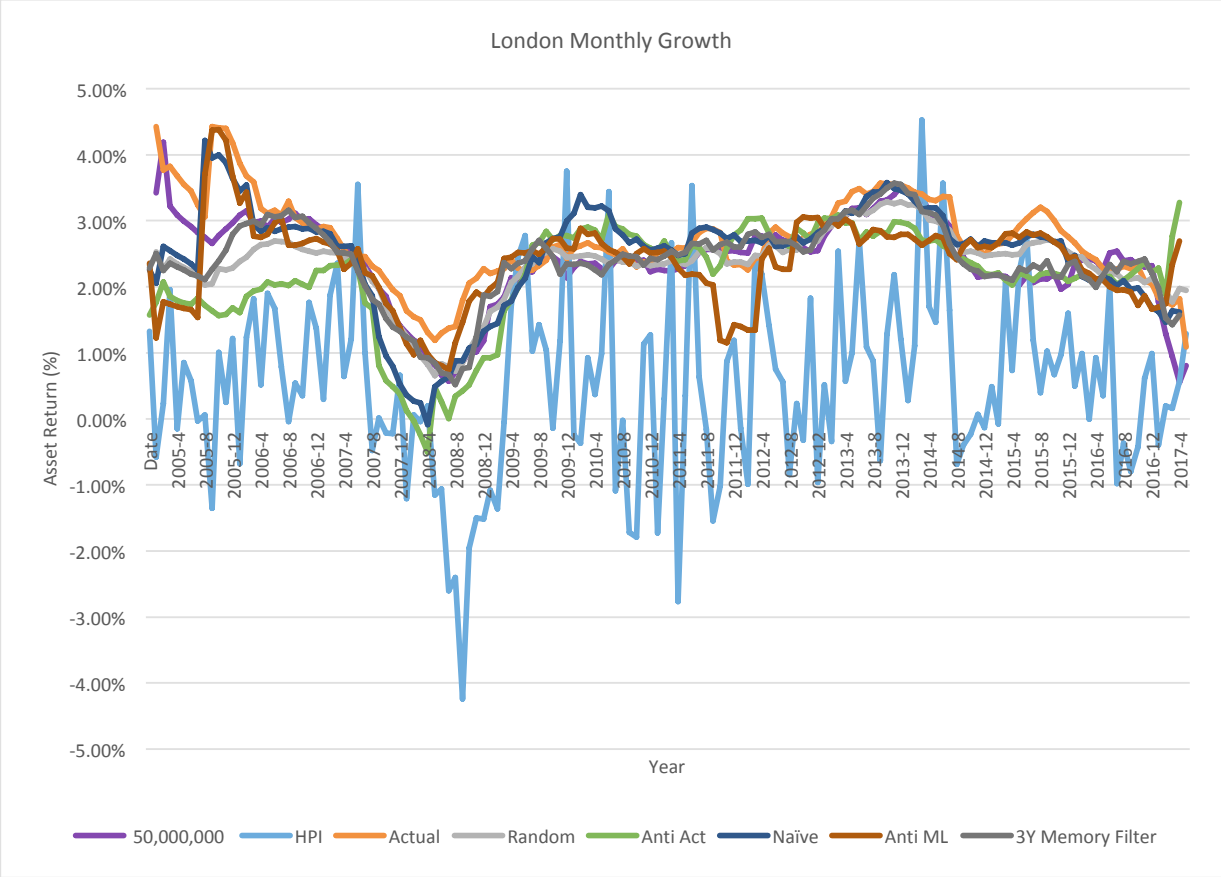


Figure 10: London Asset Returns (All)

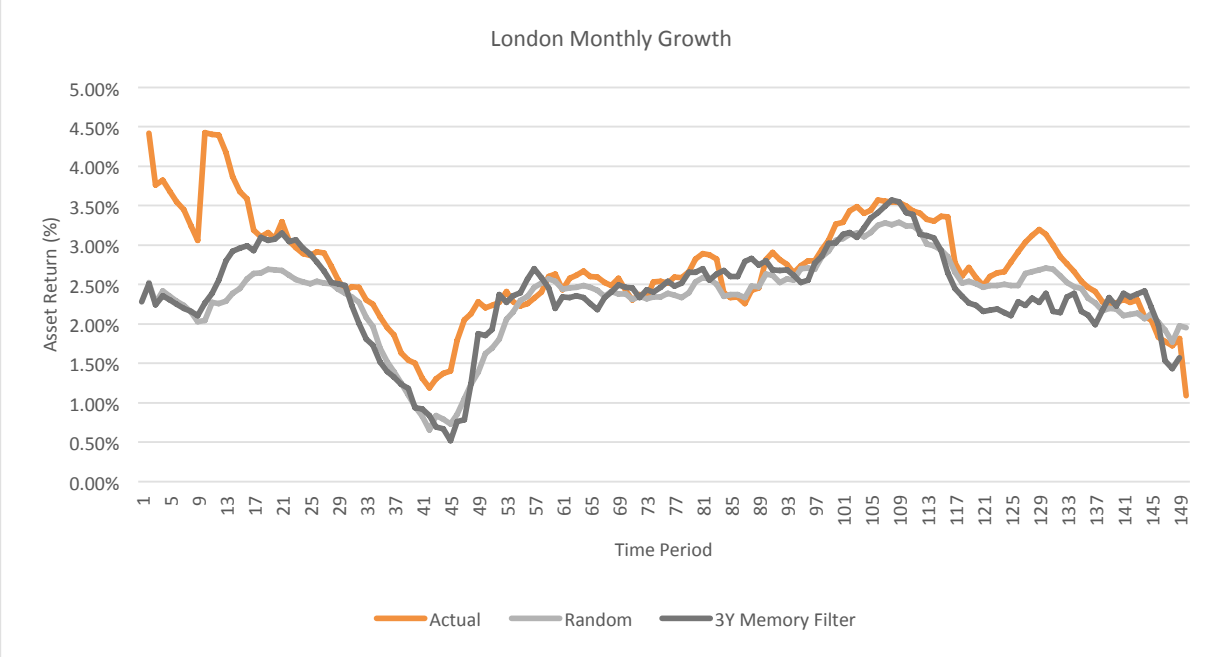


Figure 11: London Asset Returns on strategies (selected)

One can also look at the variability of the returns in the simulations performed:

Strategy	Mean	Median	Min	Max	Range (Max-Min)	Range/Mean
Random	155,500,793	155,369,116	143,134,705	171,630,669	28,495,964	18%
ML	180,953,804	181,395,859	176,257,831	185,218,094	8,960,263	5%
Anti	134,038,254	133,928,825	131,328,384	137,360,802	6,032,418	5%
Actual	194,915,724	192,811,174	189,098,690	203,043,496	13,944,806	7%
Naïve	188,105,438	187,574,041	184,318,659	193,190,782	8,872,123	5%
3Y Filter	186,762,638	187,476,138	182,169,154	190,052,853	7,883,699	4%

Table 4: Returns statistics over the simulations performed

Here we see that the various groupings of strategies are well separated regarding fund returns as their boundaries do not overlap (especially Random, Anti and Actual).

There are of course several ways that this simulation can be improved during further modelling. Several of the areas where this could be achieved without dramatically changing the approach thus far used would be:

1. Higher frequency of predictions for the Machine Learning model
2. Lag on cashflows of properties being sold (i.e. not realising cash immediately)
3. Larger/random transaction costs incorporated to account for more expensive property renovations and unexpected property events
4. Scenario modelling around buying and selling properties (i.e. buying for a greater price than market and selling at a price lower than market (although one could argue this would be smoothed out on average over all the transactions in the sample)

7.1 Brexit Test

As mentioned one issue around using the repeats sample is the presupposition of knowing the future around when a property sold (acceptable under the implementation of a strategy where buying and selling horizons are defined) and that the repeats sample may differ from the wider Land Registry. Hopefully as already discussed in this paper, these concerns should have been mitigated, but there it is worth examining what would happen if portfolio revaluation occurred on properties purchased from the Land Registry set.

As mentioned previously, the point of using the repeats transactions database is that the purchase and sale price of an asset is well defined and error free. If one uses another valuation methodology, you are obscuring the true nature of the market. For instance, an AVM that could calculate a statistical valuation of a property at any time would introduce a further layer of error into the modelling as it would be an estimated price. This is further amplified under conditions where purchases and sales deal with properties at either tail of the price distribution. An AVM estimate is intended to produce a reliable result for most of the data, and not intended to fit properties that are under or overvalued. Given a successful property purchasing strategy would involve purchasing undervalued assets, then an AVM may not be the best tool to gain accurate individual property revaluation.

In the instance though where one is examining the effects on a much larger portfolio then the use of an AVM would be acceptable as one would expect there to be some smoothing to the number of properties that would be under and overvalued through the AVM valuations. Thus, it would be possible to examine strategies applied to the wider London Land Registry data and see the effects of portfolio purchases that have not yet been sold.

It would also provide a result on the performance of the strategies over the Brexit period enabling one to see the effects of AVM revaluation, a potential structural change in the market and the use of the full Land Registry set. To this effect, the performance of the various strategies was evaluated by revaluing properties purchased using the Zoopla AVM.

Thus, the effect of Brexit was evaluated on the full set of properties available to be purchased in the Land Registry set revalued by the Zoopla AVM at the Outcode Level. The purchased portfolio was accumulated over 2016 and revalued as of August 2017. This assumes that every property available was bought and thus what is being tested is the rank orderings of the predictions.

The results of this simulation are presented in Figure 13. What is interesting to note here is the best performance of the strategies over various portfolio diversification levels. Firstly, one can see that the best performing strategy was the Anti Machine Learning predictions (i.e. the reverse ordering of the Machine Learning predictions). This suggests that a contrarian approach to market dislocation events proves most profitable. Second are the Machine Learning (filtered)/Naïve strategy, but the Naive strategies outperformance is quickly diminished below that of the Machine Learning (filtered), Machine Learning and Random/Market level. The Machine Learning (filtered) strategy dominates the Naïve at diversification of more than five Outcodes, and the Machine Learning strategy dominates both the Machine Learning (filtered) and Naïve at a diversification beyond 10 Outcodes. The Anti ML strategy becomes equal to the ML strategy at diversification of approximately 40 Outcodes.

Thus, one can see that both the Machine Learning (filtered) and the Machine Learning strategy produce above-market returns and the second-best results at a portfolio diversification of over 10 Outcodes. We see that during a dislocation event, a contrarian strategy, i.e. the Anti ML strategy produces the best returns.

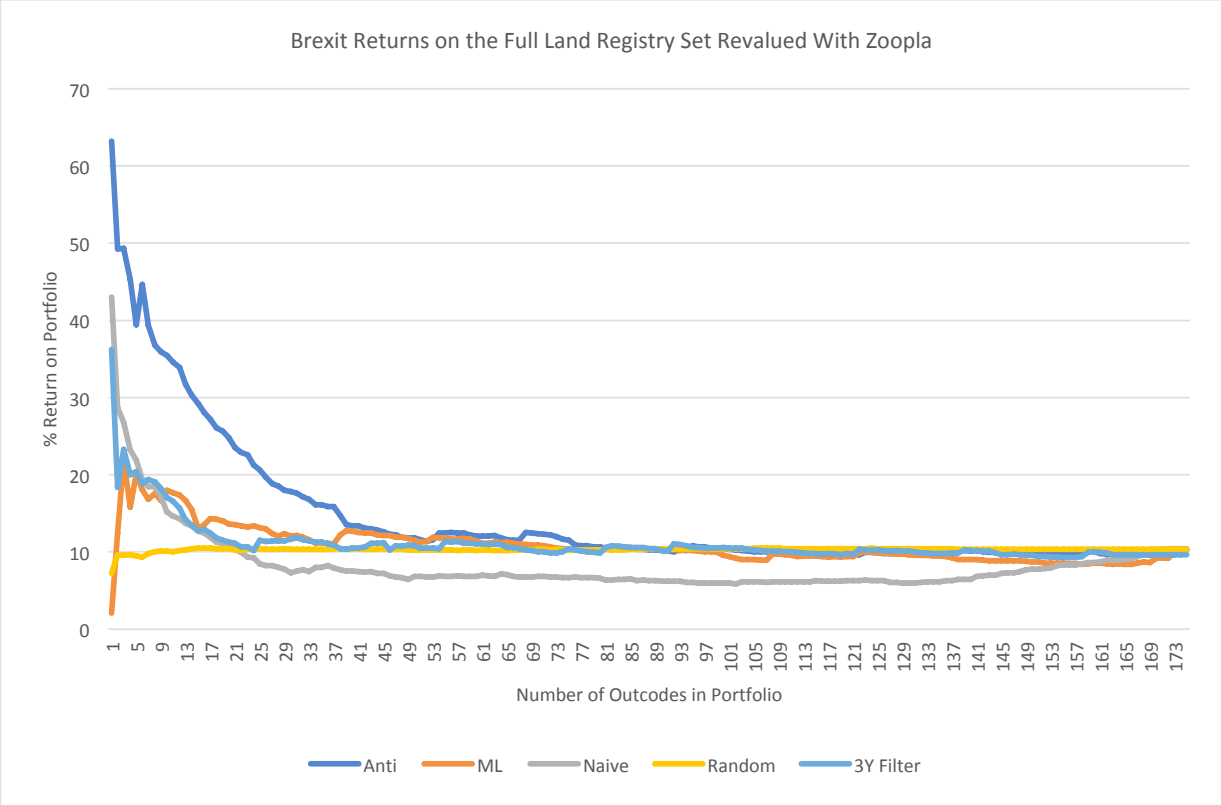


Figure 12: Brexit Returns on the Full Land Registry set revalued with Zoopla

7.2 Full Testing Of The Repeats Set

Previous evaluations of the strategies have involved applying sampling over the repeats set or looking at the full Land registry set over a specific time-period. We now look at the performance of the strategies with a simple approach evaluating their performance without sampling on the full repeats set.

In the following, we apply a simple strategy of using the predictions to purchase in one year and then sell the following year. This would mean that all repeat transactions would be purchased that for instance were bought at any point in 2005, and sold at any point in 2006, with a minimum hold period of 1 year, and maximum hold period of 2 years. This then aligns most closely with the framework for predictions that forecast over a one-year period.

We evaluate several strategies, and the results are presented in Figure 14. The returns here represent the total amount of cash returned from the total amount of cash invested into properties cumulatively spent over the entire period from 2005 to 2017. Here we see that the best performing strategy with the highest return is the Machine Learning (filtered) one that dominates all others up to approx. 50 Outcodes. After that the Naïve strategy dominates.

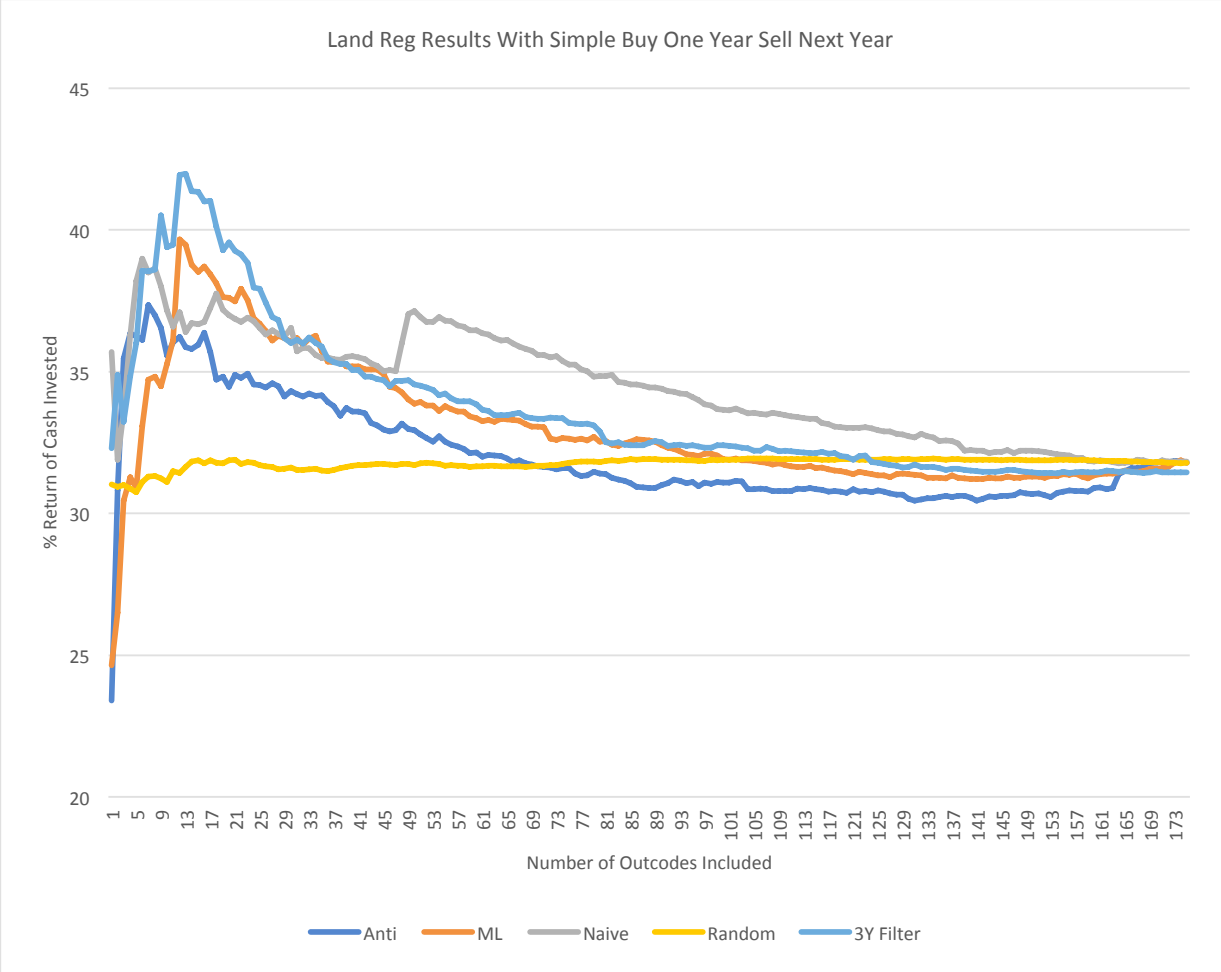


Figure 13: Various strategies on the Repeats set without sampling

8. Future Machine Learning Plans – How To Build A Better AVM

There are many ways that the model could be improved to build a better AVM, some of which have been mentioned earlier. With the current working model, improvements could come from further data cleaning and the modelling setup. Firstly, initial improvements in the data cleaning and setup could come from:

1. Increasing the frequency of prediction from annually to quarterly.
2. Division of the data into price bands (i.e. £100k-200k etc.) and modelling these bands and crossovers individually

3. Transforming the data (log transform, power transform etc.)

These immediate steps should help to improve the accuracy of the ML modelling and result in an improvement in the results in the back-testing simulation.

Improvements beyond these would require a structural change in the approach to the AVM. The first structural improvement would come from increasing the geographic resolution with which the market is modelled. The current model uses the Postcode Outcode as the modelling resolution. This could be refined to the Postcode Sector level (first part of the postcode plus the first digit of the last part). Originally the Postcode Sector level was not modelled because of the trade-off between the amount of data available and filling missing values in the time series data. In this example, merely excluding Postcode Sectors that do not have enough data points in time would result in too much loss of data.

As we go into a higher resolution of space, we do not have enough data points to provide values for all Postcode Sectors at all times. Thus, to consider this higher geographic resolution, we must look to add in additional data and a more pragmatic modelling approach. More data could be incorporated by using Energy Performance Certificate (EPC) data. This contains a variety of useful information, most importantly the size of each property in square meters. With this information, a further modelling point regarding property size can be added in and instead of looking at raw prices, one can model price per square meter (sq/m). The EPC dataset itself is not complete in time; for instance, EPC data requirements were introduced in 2007 and legally applied to all property transactions and rental transactions from 2009. The EPC database would serve to help improve the accuracy of property price predictions based on further information, but it would not solve thinness in the time series data.

Thinness is not such a problem with Outcode prices; a minimum threshold of property prices is viable each quarter to ensure that the noise in the geographic resolution is kept to a minimum. Approximately 20% of Outcodes were eliminated to reduce noise in the data. In the Postcode Sector model, many more Postcode Sectors would need to be removed (approximately half) to reduce the noisiness. This would result in the loss of considerable data. It is not a case of just using Postcode Sector data even if the transaction volume in that Postcode Sector is thin. This is because of the variation of property prices within a Postcode Sector leading to the conclusion that seemingly varying levels of price would be down to underlying directional movements, but which in fact are related to differences in the sizes of properties being sold. This, of course, is smoothed out in the Outcode resolution and at the quarterly time-period of modelling. For the Postcode Sector level, transitioning to the price per square meter approach should be a better approach because we eliminate the size differences between properties that therefore reduce the noise created because of this.

Given the reduction in noise, one can now look to model the price of sectors with incomplete time series data based on price per square meter. One could also interpolate the missing data through using nearest neighbour distances and a weighting factor, although this could also naturally come out from the machine learning modelling approach. With sector level information one could then make current time and forecasts of prices at the sector level.

After the Sector level model, one could also choose to model at the Full Postcode level. Here once again, data becomes thinner and would benefit from the inclusion of further data. At this stage, it may be worth considering modelling the actual property rather than looking at the full postcode level, but to do this effectively belies a requirement to have the entire differentiating location information on each property, i.e. its geolocation. Unfortunately, this data is not freely available and is prohibitively expensive. Geolocation information is available up to the full postcode resolution, so effectively modelling properties is modelling at the full postcode level.

So far, the primary attributes that have been used have been centred on price and location. Full modelling of properties from the area down to the point location should also include location-based data. Modelling the price per location requires fewer data at the property level, but it is worth considering from the property level to the aggregate geographic level what data would be used to improve predictions (these can also be considered as features):

Property Level		Location Level	
Location	Full Geolocation	**Transport**	Time to various Locations in London, Number of Lines, Number of Bus Stops (in various radii), Parking
Size	Size of the Property in Square Meters	**Education**	Number of Schools, Quality of Schools, Private Schools vs Free Schools, Catchment Areas, Acceptance Rates, Pupils
Features	Bedrooms, Habitable Rooms, Garage, Driveway, Parking, Facing Direction, Basement, Garden, Height, Flat roof, Solar Panels, Loft Conversion, Loft, Extensions, Heating, Energy Efficiency Rating, Double Glazing, Elevator, Floors, Bathrooms, EnSite, Planning Permissions	**Facilities**	Amount of Green Space, Parks, Width of Roads, Traffic, Pollution, Proximity to Restaurants, High Street, Supermarkets, Hospitals, GP
Condition	Age of Property, Fittings, Time Since Last Decoration, Period	**Demographics**	Age, Ethnicity, Work, Newspaper Readership, Political Affiliation, Married, Household Size, Income, number of cars, Birth Rate, Death Rate, etc
Tertiary	Asking Prices, Price Adjustments, Agent Listing and Feature Details	**Tertiary**	Area News, Facility Ratings, Social Media Commentary, Events

Table 5: Property level and location level features to incorporate into further models

Several machine learning approaches can be used depending on the resolution of prediction, but multiple levels should be leveraged with a multi-model approach operating at various resolutions, price bands and property types. For this, let us assume we model prices at the full property level (in the case of the postcode, we can just move one level lower in resolution).

1. Deep Leaning Model – Point Property – All raw data
2. Ensemble Machine Learning Model – Point Property - Enough comparables to generate enough data points in the relevant price bands
3. Ensemble Machine Learning Model – Various Resolutions – Full Postcode, Sector, Outcode

A further model can be placed on top of these models to further refine the ensemble estimate for the property price based on the above three models.

The above can be completed at both full prices and the price per sq/m. Various experimentation would be required to find the relevant price bands, number of comparables needed, dealing with missing data and relevant property type modelling. This modelling approach computes the prices of properties at the current time.

This modelling approach can then be used to forecast prices at various resolutions in time, at six months, one year, three years etc. Current time and historical current time predictions (i.e. the creation of an index)

can then be put through the same modelling approach as above with the emphasis on predicting future prices. One could just choose to model future prices instead of creating the current price AVM, but a good reliable estimate of current prices would be required as well as creating forecasts.

8.1 Deep Learning Approach

A discussion of how to implement a deep learning methodology to this problem is warranted given the rise of AVM Machine Learning companies coming onto the market.

The modelling approach used within this paper predominantly used hand designed features (feature engineering) to apply some intuitive sense to the machine learning algorithms to improve performance. With Deep Learning, one can argue that the model can be left to itself to figure out the relationships in the data.

Initial experiments performed on the previously described approaches with deep learning did not create better models because of the limited amount of data available.

When incorporating further data, deep learning should produce better results. On the individual property level, the Land Registry set contains more than 20 million transaction points, and this combined with additional features on the properties should produce a highly accurate model.

Furthermore, the general rule of training/testing splits of the data using 80/20 splitting should not be required and instead having a 99/1 split would be more appropriate given a large number of training examples.

When using deep learning, several different approaches are available. One method is to simply put all the data into the model. However, in this case, the trade-off between computing resource use/training time and accuracy will be quite evident. Training a model on all the data available may require a significant amount of computing resources and may not be as effective as focusing on techniques that utilise either transfer learning or multistage networks.

The issue that comes from using all the data is that of data mismatch. This occurs as a result of the nature of the price distributions in different areas and price bands.

It may be better to incorporate training on a wider set of data and transfer this learning to smaller geographic resolutions and price bands, thereby incorporating broader trends, but refining these for the individualist nature of areas. By this, we mean to train a neural net over London, but when applying it to smaller geographic areas, i.e. Outcodes or Sectors, the last few layers of the network are trained specifically for these geographies. Here much of the data acquired about the general market can be translated into the smaller submarkets.

Similarly, one may also transfer learning from larger geographies to smaller ones. In the creation of a multistage network, one may first choose to learn on a larger resolution (i.e. low-level features) and incorporate this into a smaller resolution. This is different to a transfer learning approach in the sense that one can use larger resolution information (i.e. area prices) rather than data points from larger resolutions (i.e. more transactions from a large resolution).

The same can also be said of different time-period resolutions. In this sense, some hand design is going into the system in the form of what data is fed and when into the neural networks.

Multitask learning can also be incorporated, such that predictions at various time points are intertwined so that the machine learning considers what the relationships are between different prediction time features and area resolutions. This may be better than using multiple individual models.

The last point to briefly touch on is that of the limitations regarding accuracy for any modelling approaches. Usually, the Bayes Error is deemed to be upper limit at which accuracy can tend to, and for property price evaluation regarding current prices, this tends to be around 5%. This is usually because prices tend to be anchored at the initial selling price and then purchase offers tend to be accepted at less than ten percent from this estimate, on average at around five percent from the initial price. This is the element of human negotiations that go into unobjectively pricing property, and perhaps accuracy for current time-period human error may be close to the Bayes error.

In the case of forecasting, human errors are much larger as much evidence pertains to when property researchers have dramatically and on many occasions incorrectly forecasted what will happen in the market one to three years into the future. In this time domain, the scope to be better than human forecasts may be much more significant.

9. Application of the Model in Real Estate Markets

With a current price and forecasting price model complete, one would need to understand how to apply this model to take advantage of the forecasts. The basic premise would be as follows:

1. Rank forecasts based on property or area over whatever time periods you are interested in acquiring
2. As properties come onto the market, target those areas with the highest expected growth and use the current price AVM to acquire undervalued properties.

Ideally, in real time, the setup would enable all new properties that are coming on the market to be evaluated for their acquisition potential based on capital appreciation, development potential and income potential. This would be done by evaluating property listing sites. Separately one would evaluate auction sites to create optimal bids for undervalued properties. Given the nature of relationships in Real Estate, one would also require a team of trading agents to be developing relationships in areas of high growth to find deals that are not advertised on the wider market. Another simpler approach is to arbitrage mispricings from Developers whose cost of development and sale values of properties can be quite wide. In these situations, one can buy in bulk and achieve significant discounts and operationally New Build properties are regarded as better buys for their lack of maintenance costs (which can be a significant cost on older properties).

10. Case Study On Properties Bought In One Month

To better understand what is going on with actual properties bought through the strategies applied in the market, a summary of transactions that occur in the simulation is presented below over randomly selected months and can be found in Appendix 3. We look at December 2005, June 2008, February 2011 and September 2014. The returns during those months versus the House Price Index are presented below:

Date	Portfolio Buy	Portfolio Sell	Portfolio Return	Portfolio Return (after 10% transaction cost)	HPI Buy	HPI Sell	HPI Return	HPI Return (after 10% transaction cost)
Dec-05	£ 1,467,500	£ 2,068,495	41.0%	28.1%	£ 237,127	£ 274,191	15.6%	5.1%
Jun-08	£ 1,524,000	£ 1,655,000	8.6%	-1.28%	£ 291,896	£ 273,328	-6.4%	-14.9%
Feb-11	£ 1,870,000	£ 2,267,500	21.3%	10.2%	£ 285,549	£ 297,924	4.3%	-5.2%
Sep-14	£ 5,739,500	£ 7,025,000	22.4%	11.3%	£ 411,840	£ 470,317	14.2%	3.8%

Table 6: Portfolio purchases by month case study

We see from studying these properties that the 10% transaction cost applied to these properties actually represents a significant hurdle if one considers a trading strategy. In this case, properties that are selected must somehow surpass this initial hurdle and can do so only in the case of some form of development or renovation being considered. By this one means to buy older properties and to refurbish them.

Of course, in this case, the true return would be affected by the costs of refurbishment, which can be quite expensive (£50,000 - £100,000). But these costs when incurred at a portfolio level can experience economies of scale and be significantly reduced, and one must also consider that the optimal strategy is to buy and hold for a number of years, rather than incur large transaction costs through a higher frequency of trading.

When purchasing properties in the market, one can assume that a discount would be obtained in the form of being an all-cash buyer (i.e. from the ability to rapidly transact) and through buying off-plan from developers (when close to completion), counteracting the burdensome effects of the costs of refurbishment experienced in this model, thus making these results comparable with results achievable in the market.

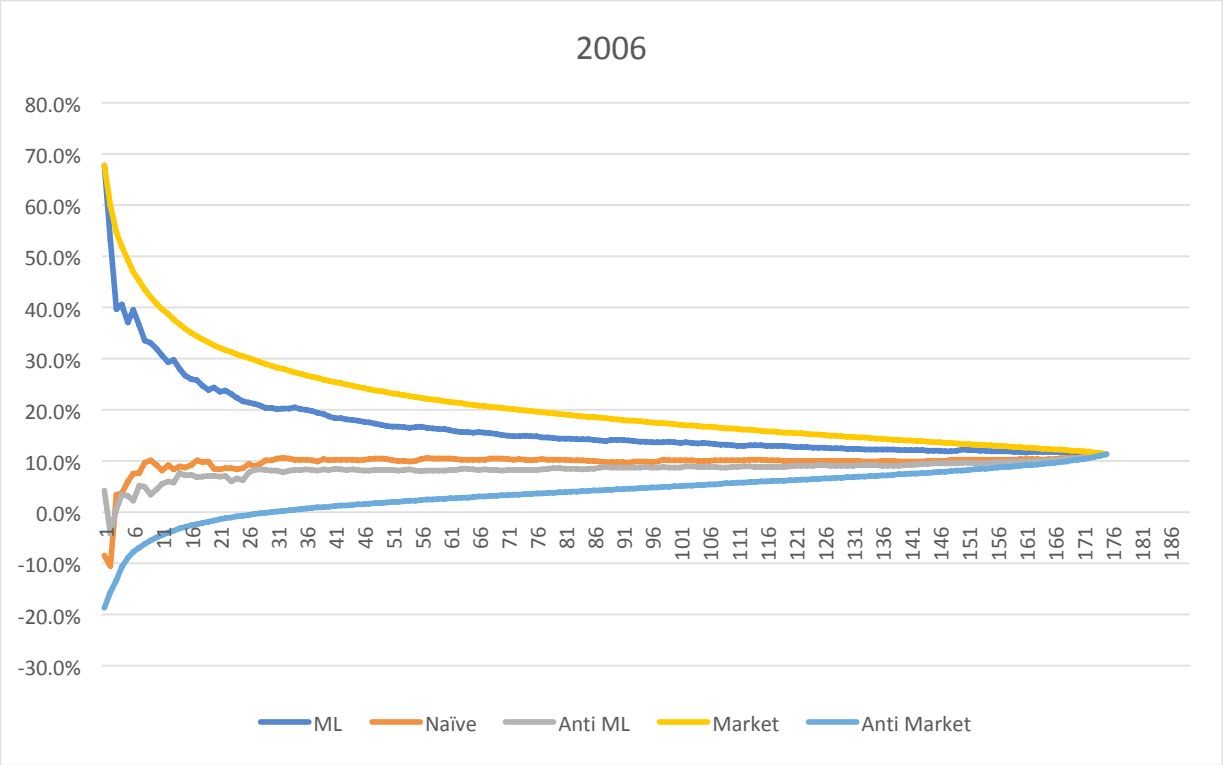
11. Acknowledgements

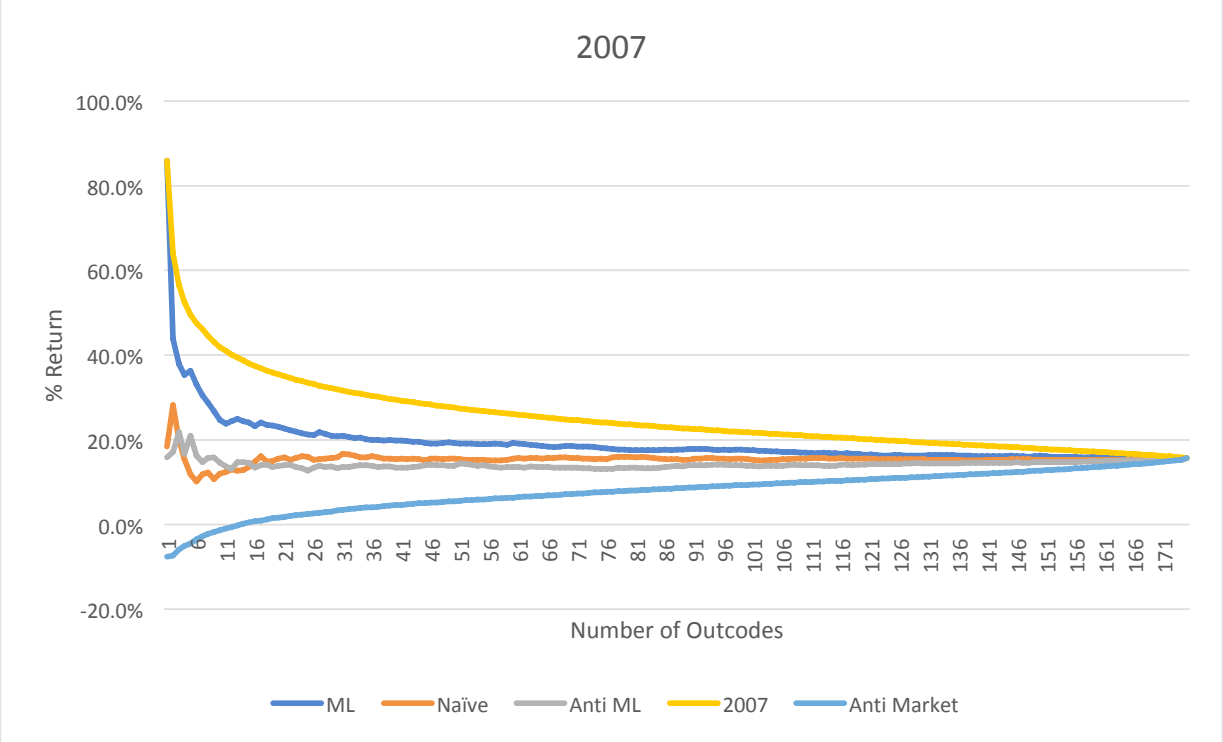
This piece of work was primarily put together by Nikhil Vadgama and Hamid Seyedsayamdost with the assistance and advice of many individuals, most notably Nicholas Russell who was a part of the creation of several parts of this work. Dr Alastair Moore, Niall Roche and Daniel Fozzati from UCL were very kind to allow the use of computing resources and to discuss various aspects of the developments of the models and business strategy. Thanks, must also go to Dr Lee Mollins who has imparted a great deal of knowledge into how AVMs and the AVM business works.

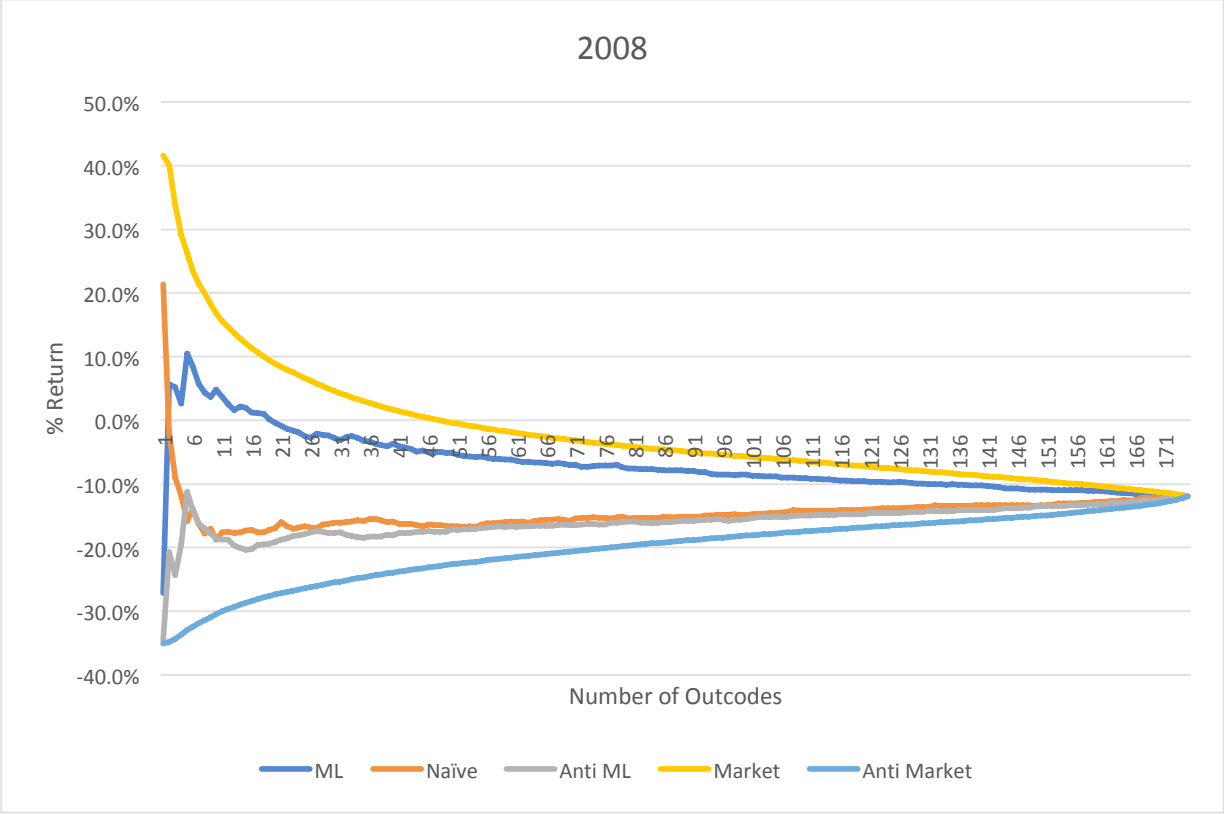
This work has primarily been produced for a business called Propcoin that is ultimately seeking to introduce the concept of equity financing for Residential Real Estate; to make home ownership affordable once again.

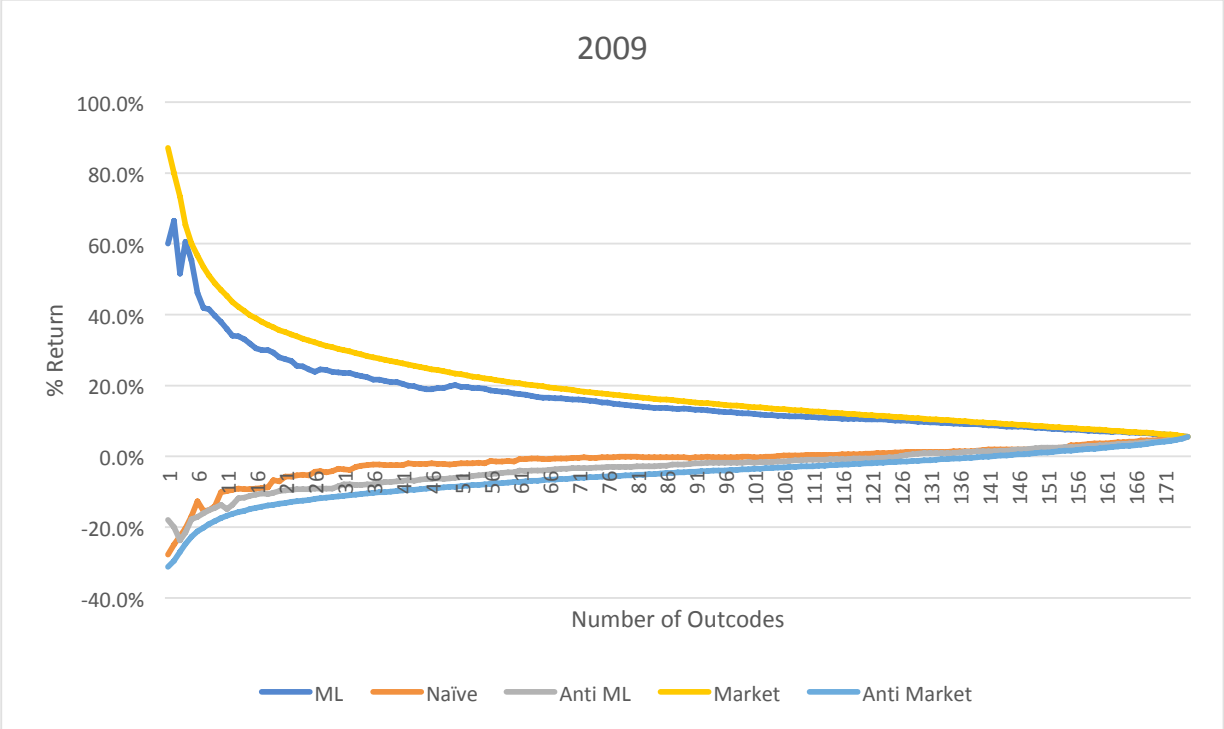
12. Appendix

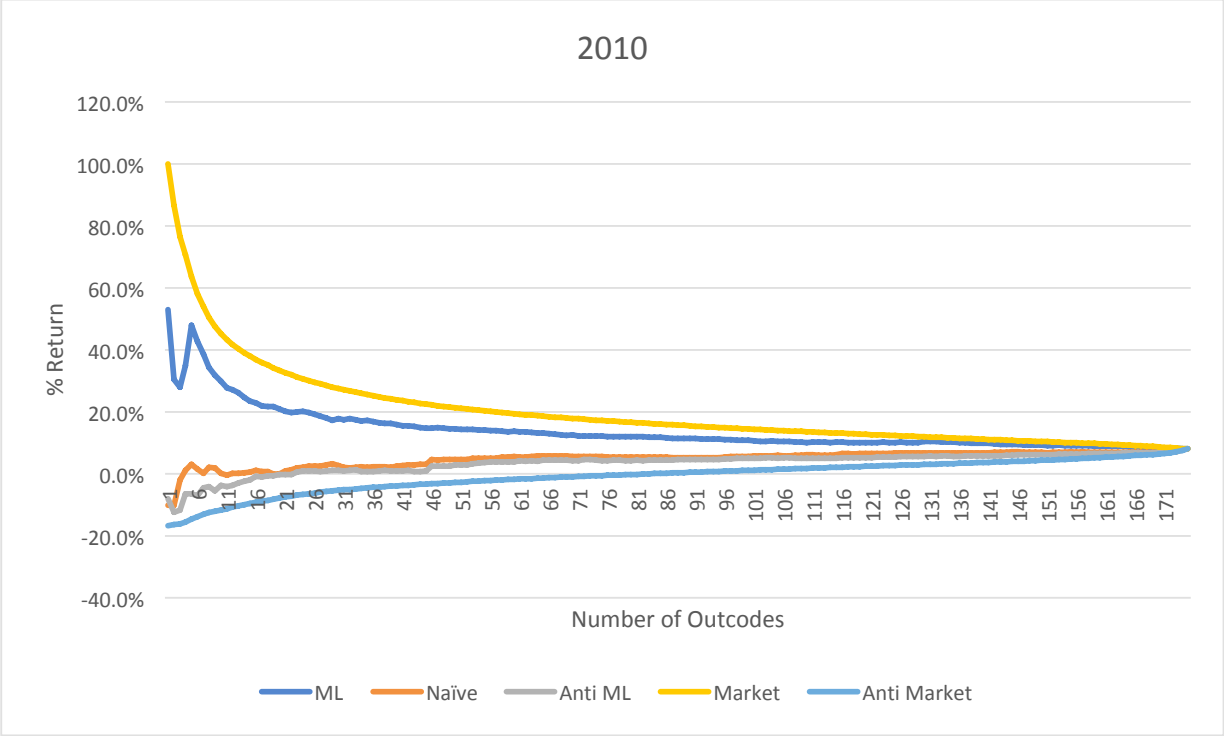
12.1 Appendix 1 – ML Results by Year

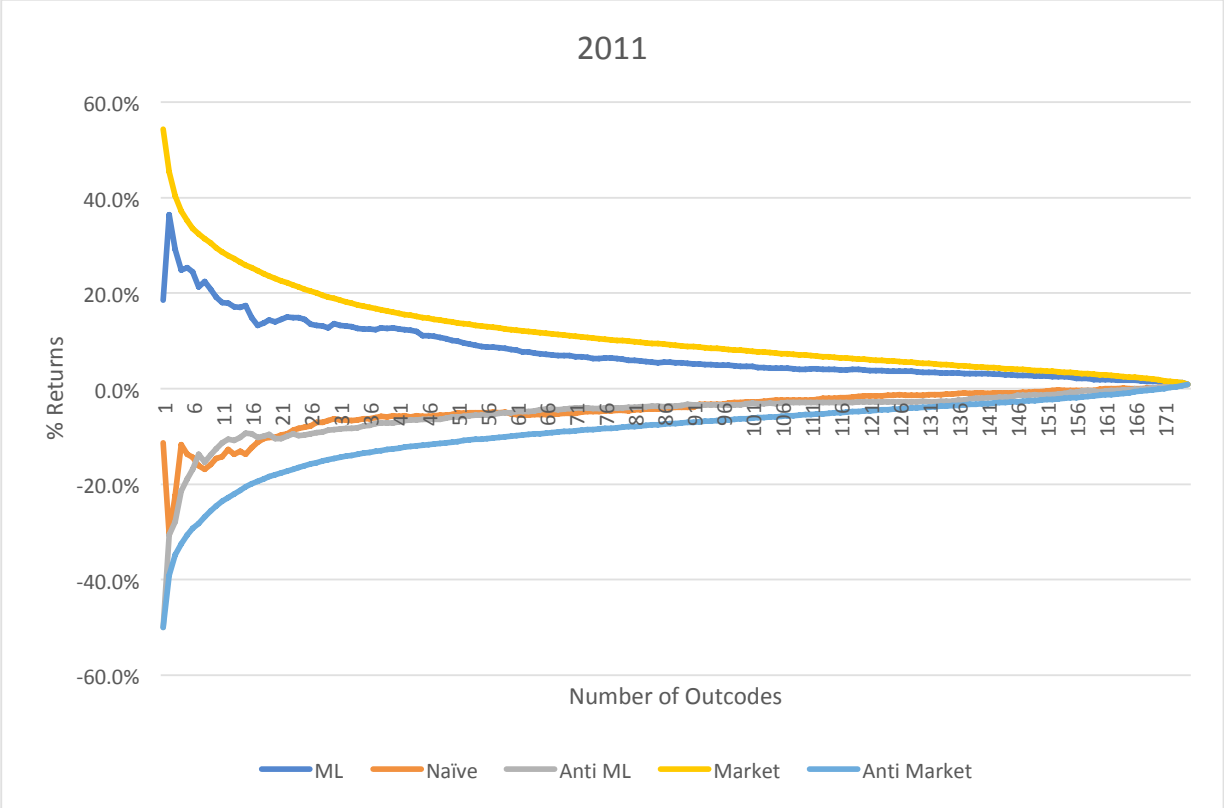




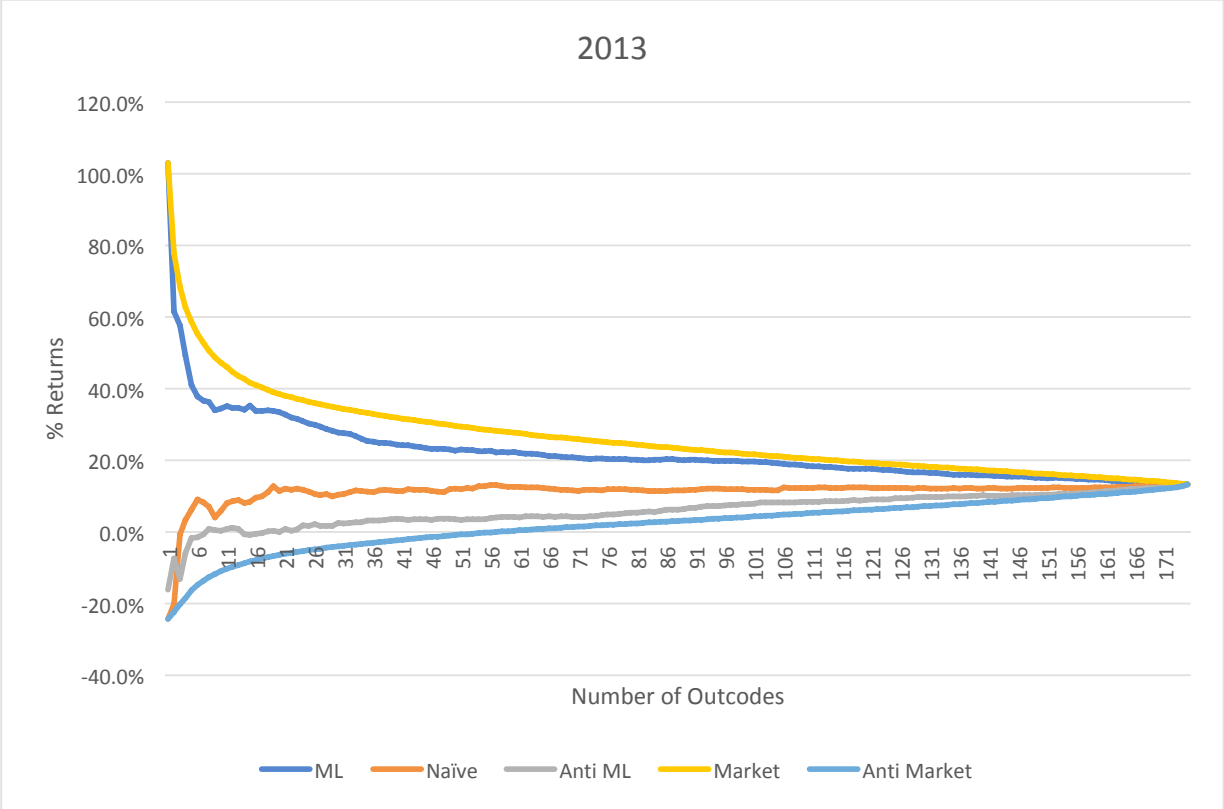

2008
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
-10.0%
-20.0%
-30.0%
-40.0%
% Return
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
101
106
111
116
121
126
131
136
141
146
151
156
161
166
171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market

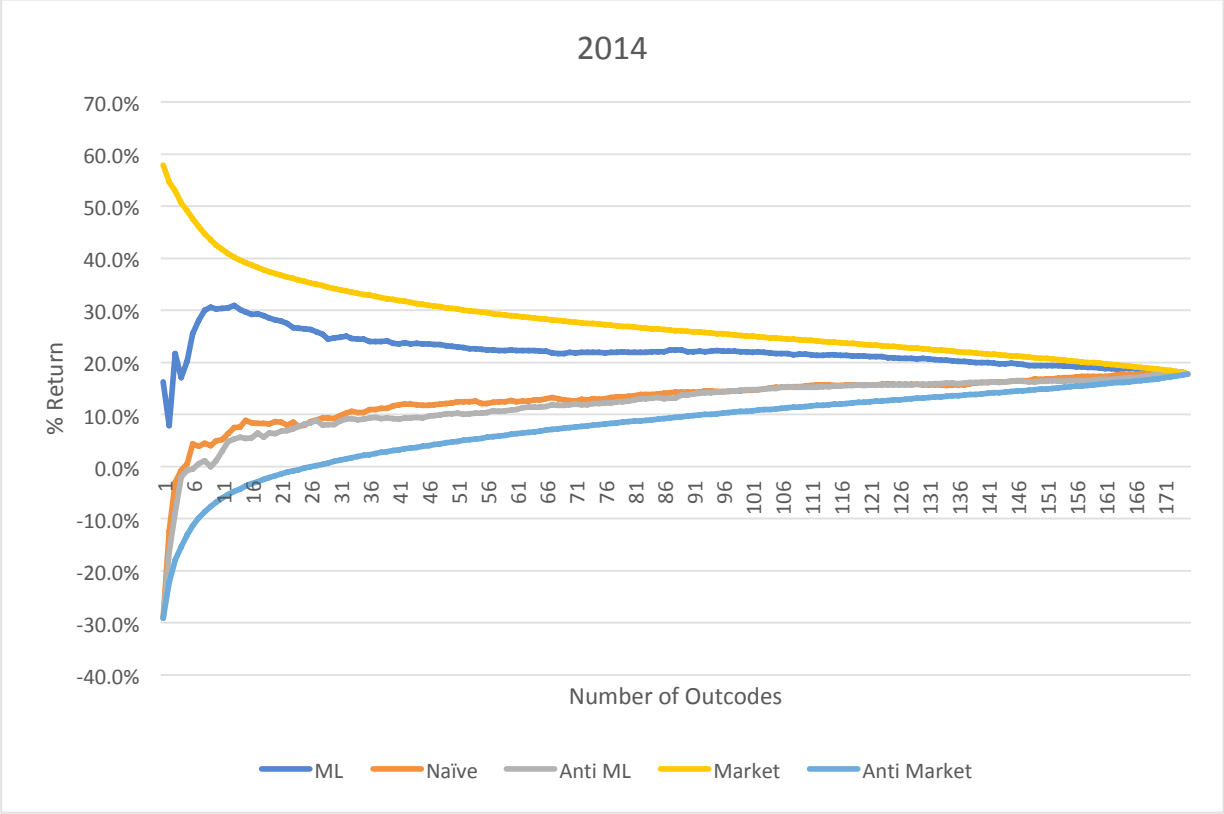

2009
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
% Return
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
101
106
111
116
121
126
131
136
141
146
151
156
161
166
171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market

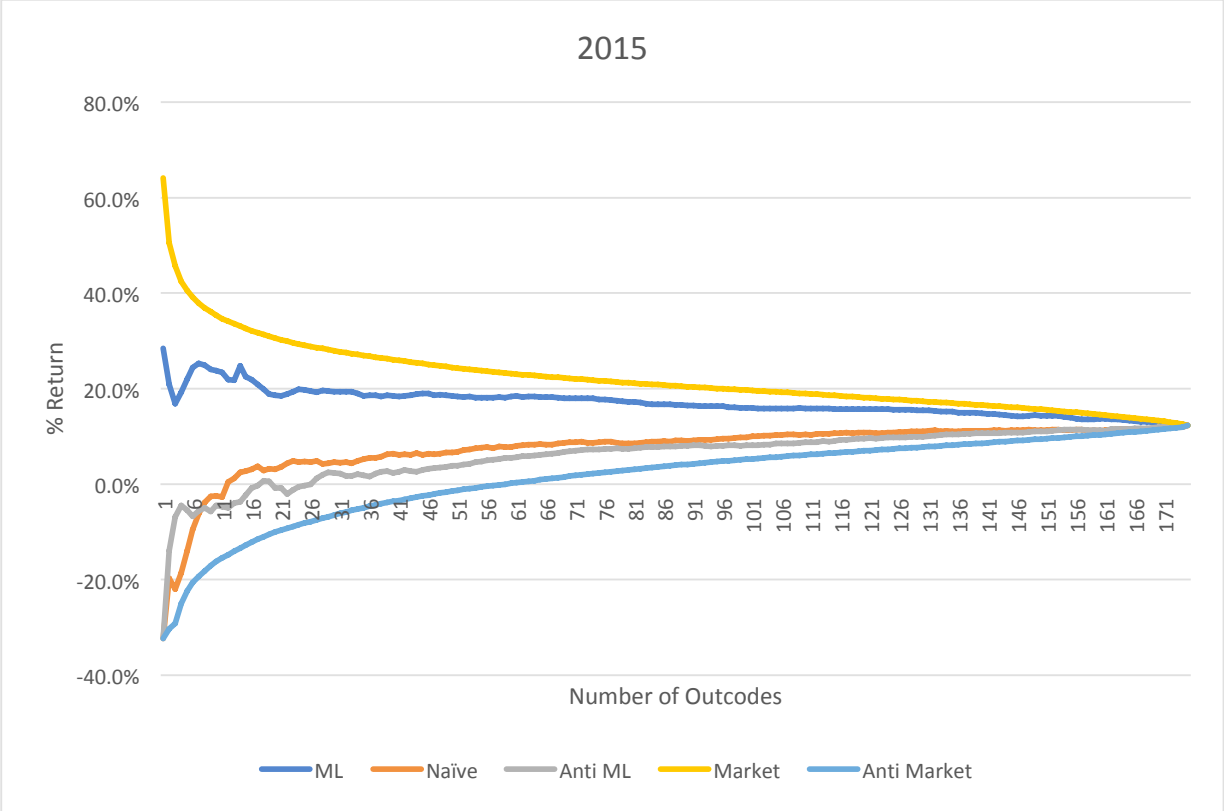

2010
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
% Return
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market

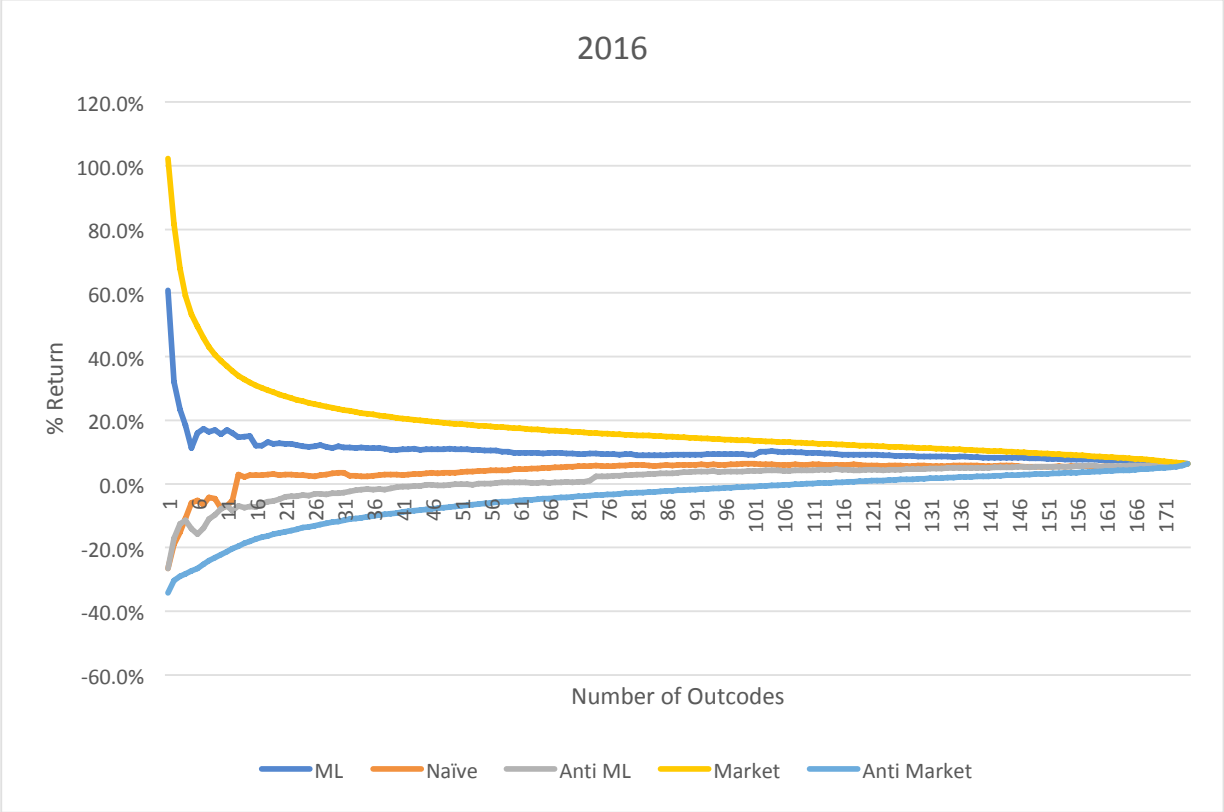

2011
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
-60.0%
% Returns
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market


2012
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
% Returns
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market


2013
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
% Returns
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market


2014
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
-10.0%
-20.0%
-30.0%
-40.0%
% Return
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market


2015
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
% Return
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market


2016
120.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
-20.0%
-40.0%
-60.0%
% Return
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
101
106
111
116
121
126
131
136
141
146
151
156
161
166
171
Number of Outcodes
ML
Naïve
Anti ML
Market
Anti Market

12.2 Appendix 2 – Price Distributions By Year

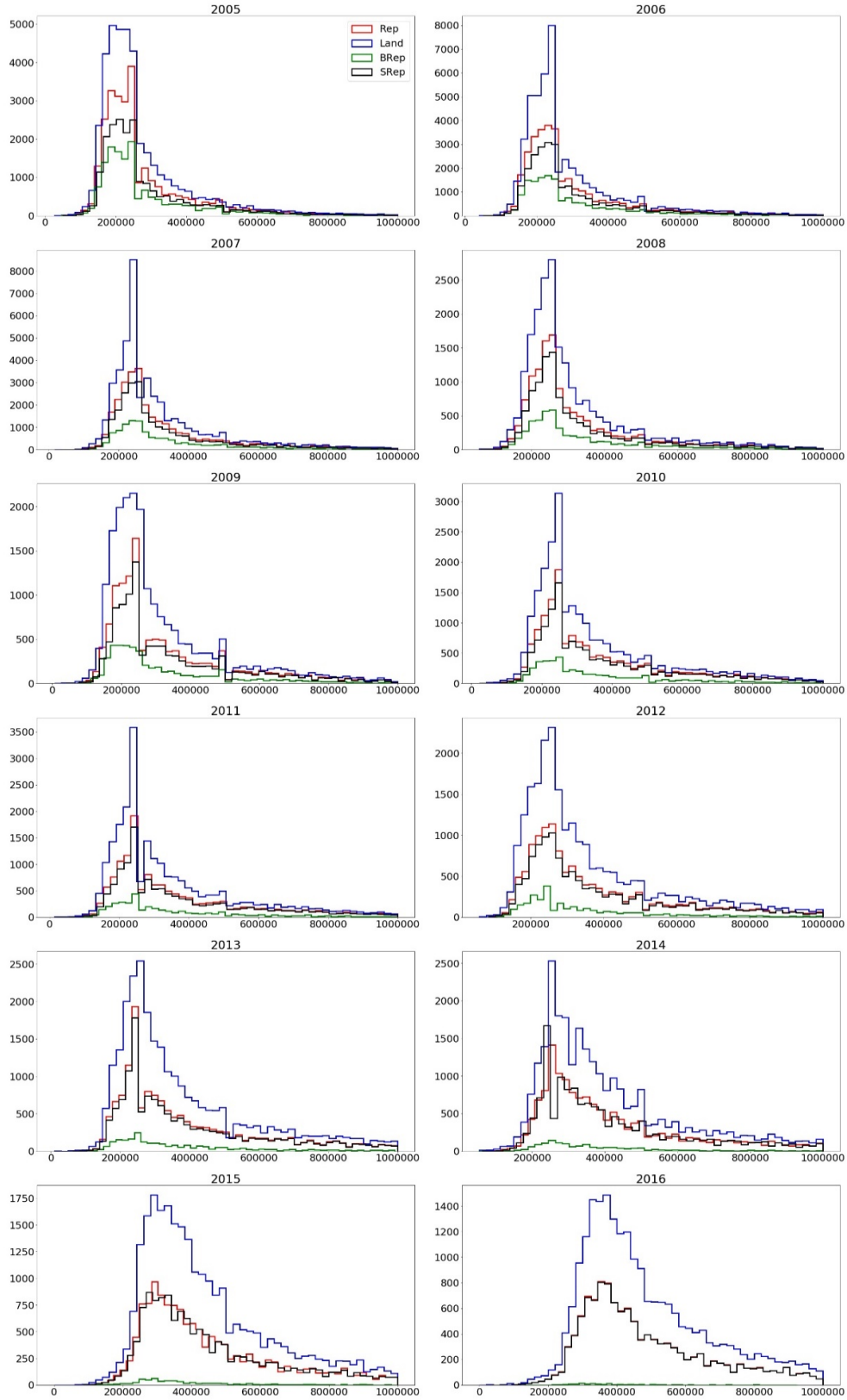


Figure 14: Histograms of transaction price on the various datasets by year (X-axis: Price; Y-axis: Frequency