PREDICTING HOUSE PRICES

AN ANALYSIS OF GREATER LONDON TERRACED PROPERTIES

Raj Gupta, 23/01/2018

2 INTRODUCTION

- "Bubble" in central London house prices.
- Prices drop off significantly as we move away from central London.
- Ought to be a clear relationship between house price and location.
  - Can this be dimensioned and predicted?
- Focus on location (postcode) rather than individual house
  - smooths out idiosyncratic features and renders the data homogenous.
- Are there any other determinants?

3 DATA SET-UP

- Data on prices paid obtained as csv files from HM Land Registry's website.
  - Yearly files going back to 2004.
  - Provides data on all property transactions in the UK by transaction date, price paid, house type, postcode, street address, duration of lease etc.
- Data on geospatial co-ordinates of postcodes.
  - Latitude and longitude
  - Needed to calculate distance of property (postcode) from Central London.

4 TRAINING AND TEST SETS

- Used to build models and test predictions.
- Choose between 50% - 75% of the dataset as training set.
- Training set further split into a validation set between 15%-25% to compare models.
- Dataset divided into temporal subsets of previous {2, 3, 5, 10} years.
- Are the predictions stable over split and time?

5 FEATURE SET

- Greater London properties only.
- Transaction date,
- House type i.e. terraced, detached, semi-detached or flat
- Distance from a reference point (Central London)
- New build or old
- Outright transaction or buy-to-let, repo etc.
- Freehold or leasehold.

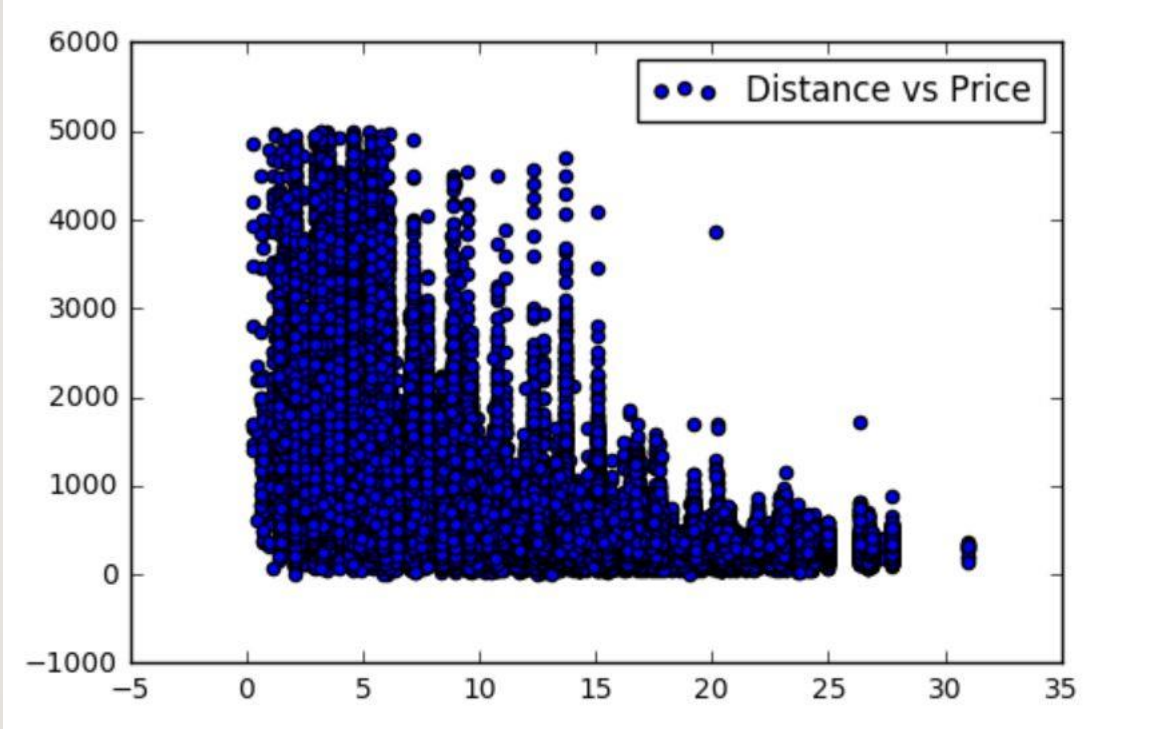
6 FEATURE ENGINEERING

- "Terraced" category is most homogenous.
- Streamline location feature by using out-codes (the first part of the postcode).
- Out-codes converted into distance (km) from spatial co-ordinates.
- Central London postcode WC2N used as the reference point.
- Time-trend introduced to capture the general appreciation in property prices.
- The other features: build type, transaction type, duration type are not important as predictors of house prices.

7 DISTANCE IS KEY

Exponential damping relationship between price paid and distance. We thus transformed the target to log values.

Steep decline in property prices as distance increases -> an indicator of Central London "bubble".



8 MODEL SELECTION – LINEAR MODELS

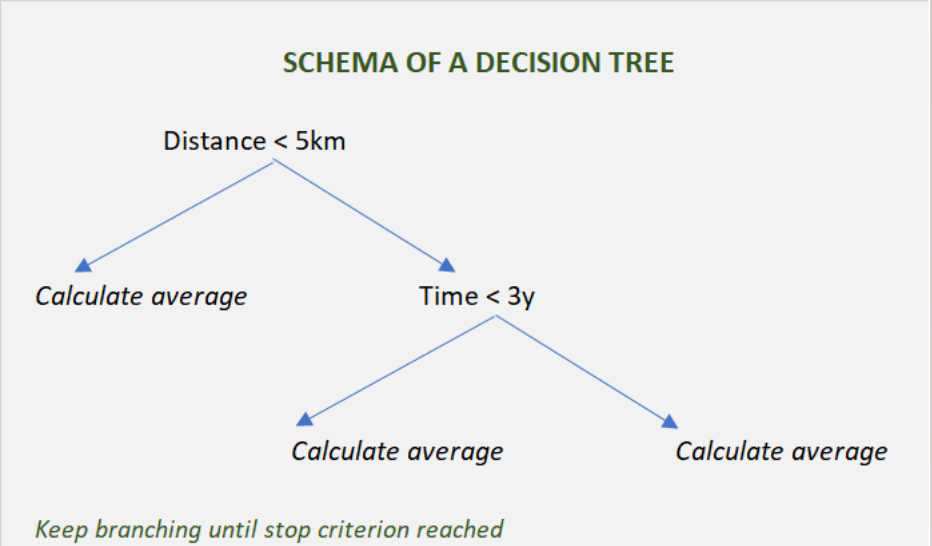
- Benchmark model: Linear Regression (LR)
- Key features: Distance and Time trend.
  - Build type, transaction type etc. had hardly any explanatory value.
- Model performance captured by R-squared
  - Measures the proportion of the total variation in prices "explained" by the key features
  - Number usually between 0 and 1
  - Higher the R-squared, better the performance.
  - R-squared roughly 45% for LR and more general models (Lasso, Ridge and ElasticNet)

9 DECISION TREES – A PRIMER

NODES AND BRANCHES

- Start with a feature and divide the dataset into two regions.
- Branch out from one of the nodes and obtain three regions.
- Continue to the specified "depth"
- Repeat until terminal nodes ("leaves") have minimum variance.

SCHEMA



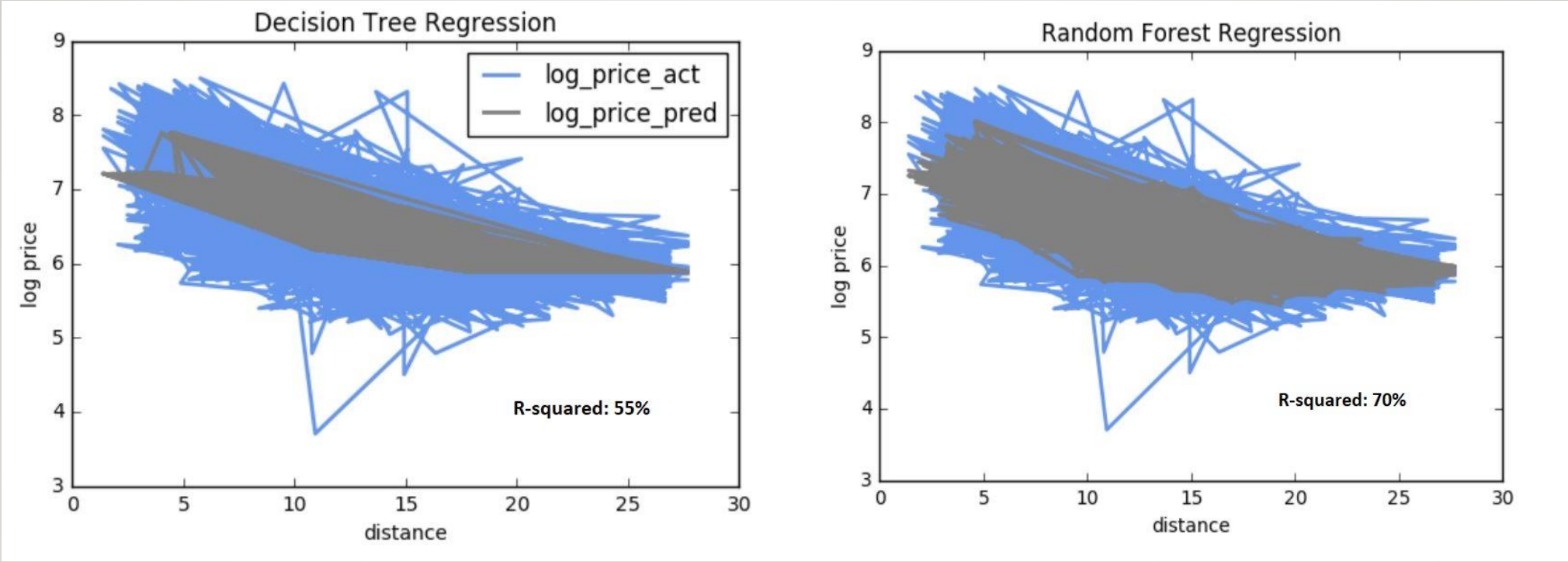
10 ADVANTAGES OF DECISION TREES

- Non-parametric approach.
- Can handle non-linearity.
- Problem of instability does not apply
  - Data set is remarkably stable
- Performs better than Linear models:
  - R-squared (percentage of total variation explained) around 55%.

11 ENSEMBLES OF TREES

- Ensembles enhance tree model by repeated sampling.
- Construct many decision trees using random draws from the dataset.
- Popular model: "Random Forest"
- Idea: combining trees likely to provide a better quality fit.
- Result: Significant improvement vs Decision Trees.
  - R-squared was around 70%.

12 PERFORMANCE: TREES VS ENSEMBLES



13 CONCLUSIONS

- Property prices depend significantly on distance from Central London.
- Fit quality is stable over time.
- Tree models perform better than linear models, ensemble models even better.
- (Next step) Build a predictive model that uses lagged price variables to predict future out-code prices. We plan to use EPC data to generate more features.
- (Extension) Predict rank-ordering of out-codes by price change using Learning to Rank models.