We List Houses

By: Sean Carver & Ngoc Tran
July 31, 2019

Introduction

Are you:

- Upgrading to a bigger home?
- Moving somewhere else?
- Getting a divorce?
- ...

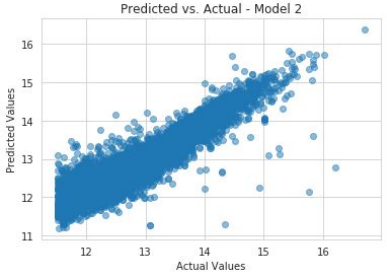
Feeling stressed? Don’t worry! We can help:

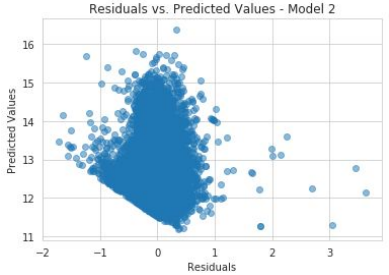
- Price your house
- Pick the best time to sell

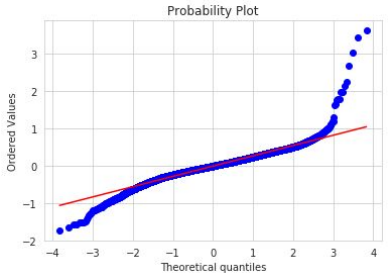
We know what we’re doing! Our Machine Learning models were built on well processed data:

- 54,425 D.C. residential properties
- 325 features
  - Location
  - Living area
  - Number of rooms
  - …
- Target variable
  - Price

Linear Regression





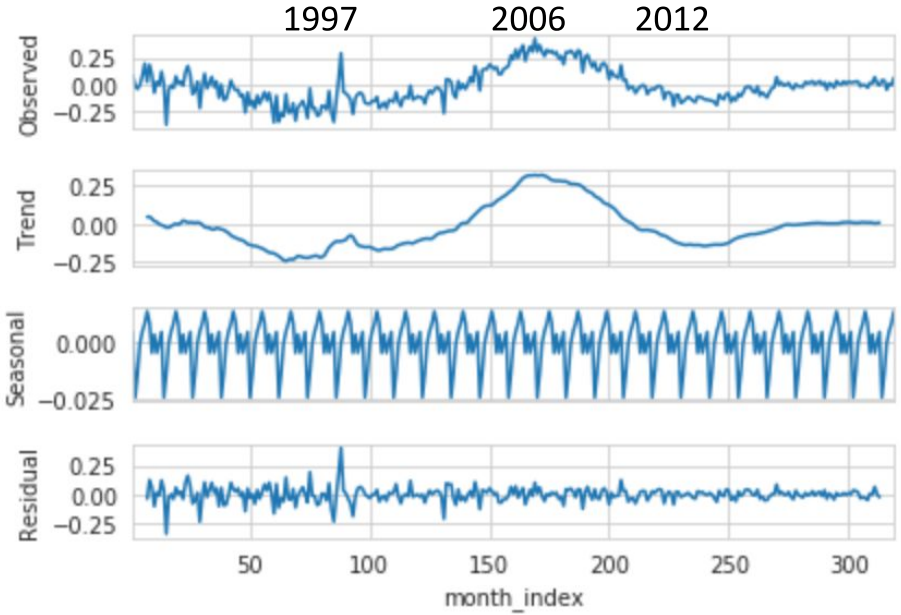


- Linear regression gives client a fair market value for his/her house.
- Our best multivariate linear regression model:
  - Data:
    - All features
    - log(Price)
  - Performance:
    - Non-overfitting
      - Adj R-squared on training set: 0.8573
      - Adj R-squared on test set: 0.8530
    - Root mean squared error: 0.28

Logistic Regression

- The client may already have a sale price in mind (e.g. $1M).
- Logistic regression answers the question: what is the probability of selling the house at the target value?
- Model Selection: Including sale date is a better model than not including it.
    - Delta_AIC = 827.2036511688116.
- Model Selection: Using 46 dummy variables (for number of bedrooms, number of full bathrooms, number of half-bathrooms, number of extra rooms) is a better model than using 4 quantitative variables for these quantities.
    - Delta_AIC = 1960.7685314010923 (despite penalty for extra parameters).

Seasonal Decomposition



Next Steps for Future Improvements

- Reduce multicollinearity for model interpretation
- Perform further feature engineering and model selection
- Use more advanced algorithms for better performance
- Build a time-series model for extrapolated predictions

Conclusions

- Linear regression gives a client a fair market value for his/her house. It looks like error assumptions are violated for linear models. Further EDA will need to be performed for better feature engineering. More advanced Machine Learning algorithms can produce better results.
- Logistic regression gives a client a probability of selling his/her house at a target value. Model selection was performed but more is necessary to obtain the best model.
- Dummy variables for numbers and types of rooms work tremendously better than many fewer quantitative variables, despite the penalty for additional parameters. We attribute this behavior to nonlinear effects.
- Seasonal decomposition works as expected showing trend and seasonality in residuals.

Questions?

THANK YOU!

Sources

- https://www.kaggle.com/christophercorrea/dc-residential-properties
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.seasonal_decompose.html
- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html