Predictive Analytics Case Study- The Effect of Covid-19 on the Real Estate Market

DSC630 Moshe Burnstein

Introduction

The common assumption is that a pandemic depresses the property value of apartments, as was painfully displayed during the 2007 crash. Andrius Grybauskis, Vaida Pilinkiene, and Alina Stundziene published an article in Springer Open on August 3, 2021, called "Predictive Analytics Using Big Data for the Real Estate Market During the Covid-19 Pandemic" . This study leverages Big Data to prove what really drives apartment prices. One particular attribute which has vexed real estate experts is the time-on-the-market variable (TOM). Some studies have shown a positive correlation between TOM and price, some have shown negative correlation, and some have shown no appreciable correlation. These issues are crucial for real estate investors as well as for buyers and sellers. If the real estate market does not depreciate so significantly during times of crisis, one would need not hold off on selling his house then out of fear of taking a loss. Conversely, the buyer would be aware of the true value and not make inappropriately low offers. The economy in general would have more confidence in the real estate market if one could prove that such catastrophes do not wreak havoc on prices.

Methods and Results

The data for this study was gathered by a Python web scraping program using the BeautifulSoup and Selenium packages. This algorithm collected data of all the variables for the apartment listings in Vilnius, Lithuania. The data were collected separately for the four months May, June, July and August. These months represent times of "opening up" and "shutting down" in response to the increasing and decreasing of the infection rates respectively. The program actually scraped 18,992 apartment listings.

The algorithm produced 16 features. These features include city zone of the apartment listing, address, listing price, number of rooms, apartment size, the number floor where the apartment is situated, the number of floors of the apartment, the change in the list price, the year built, the distance to the shop. The distance to the kindergarten, the distance to the school, the building type of the apartment, the heating type, the vacancy, and the price change date. Data processing converted the heating type variable which had 40 levels into a more manageable 13 levels. They used a targeted algorithm which scraped exactly what they wanted so they did not have as much preparation to do with the data.

They converted the target variable of price change for each of the four months into dummy variables. They did the same for the location variable. They opted for target encoding for the heating and building type because they were wary of adding so much more noise and dimensionality to the data. The attributes rooms, number of floors in the building, and number floor of the apartment were all encoded ordinally. They state that they did this to preserve rank. It also keeps the dimensionality of the data down.

The researchers first split the data into 70% training and 30% testing sets respectively. They chose to score their models on the training set also, even though doing so brings the risk of overfitting the data and not producing a true model which can confidently predict on new as yet unknown data. They did this to ensure that they would be able to use Shap values later to evaluate the features' importance.

Scott M. Lundberg developed **SH**apley **A**dditive ex**P**lanations (Casas, P., 2019). SHAP values are the calculations of the difference in how the model predicts when including the features and when excluding them (Tseng, G., 2018). The SHAP library further optimizes its calculations using all combinations of features by using the individual model's structures (Tseng, G., 2018). This is a boon for interpreting these previously black-box complex algorithms because one can discern how it works (Lundberg, S. M., & Lee, S.-I. 1970).

They trained 15 models on the data of the four months. These models span the gamut of classification algorithms. The researchers chose models which represent all the methods of classification. They would run all of them, choose the highest scoring model, and tune it to get the best performance out of the best performer. These models included CatBoost Classifier, Extreme Gradient Boosting, Light Gradient Boosting Machine, Random Forest Classifier, Extra Trees Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis, Logistic Regression, Ridge Classifier, Naïve Bayes, Ada Boost Classifier, K-Neighbors Classifier, Decision Tree Classifier, Quadratic Discriminant Analysis, and SVM-Linear Kernel. During stratified cross-validation they employed the SMOTE technique on the training set to offset the imbalance of the data. The Synthetic Minority Over-sampling TEchnique (SMOTE) generates new samples to supplement the minority class (Elreedy, D., & Atiya, A. F., 2019). It leverages K-Nearest Neighbors and uses Euclidean distance to create these synthetic datapoints (Elreedy, D., & Atiya, A. F., 2019).

The researchers scored the models on seven metrics: accuracy, area-under-the-curve, recall, precision, F1-score, Kappa value, and Matthews Correlation Coefficient. They leaned most heavily on accuracy, precision, and F1-score to account for the preponderance of negatives in the target. They set a random seed to ensure reproducibility.

The XGB model proved most robust based on their metrics. The tuning stage included running the stratified cross-validation with SMOTE. The hyperparameter tuning was done using a grid search. They then applied the SHAP values to add interpretability to this complex model. SHAP values represent the magnitude of the predictive power.

This study showed that first and foremost most properties of both rentals and sales did not depreciate over the May to August period. Only 17.2% of rentals and 10.7% of sales properties saw prices decrease. The price decrease for rentals occurred after an average of 23 days on the market, and after 63 days for sales. The magnitude of the price drops decreased from the first to the fourth month. This corresponds to the peak of infections reported in May and the lower numbers of infections in the successive months, to the end of the quarantine on July 16. The average TOM rose from 21 to 24 days for rentals and from 31 to 45 days time-on-the-market for sales. According to Lazear (1984) longer TOM indicates economic downturn because people are reticent to put their money into new rentals or sales. Yinger (1981) posits that longer TOM is indicative that people wish to maximize their profit and are willing to wait until they receive their higher asking price. Many people assume that people move away from cities in response to pandemics. The data did not show that people were relocating to sparser populated areas, dispelling this notion.

Based on the SHAP scores that the researchers used, the TOM variable proved to be the most important feature correlating to a price change, over these four months, for rentals and for sales. For sales the second and third most important features were year and initial price, while for rentals only the initial price variable made any meaningful contribution, after TOM. In a global sense the researchers have confirmed He et al.'s hypothesis that TOM compared with price is not a linear relationship, but rather more closely resembles an inverted 'U'. The initial TOM period suggests a rise in price and a later period indicates a drop, albeit not as defined as the original rise.

The greatest takeaway from this exhaustive study is that the TOM variable is the one which has the greatest predictive power. One must take a nuanced view because TOM variable is positively correlated with price in some periods, and negatively correlated with price in other periods. Buyers, renters, sellers, and investors must all realize that while TOM will indicate the 'real' price of the property, the specific time period of the TOM is the true driver. The researchers assert that governments and entrepreneurs would do wisely to follow the TOM 'indicator' and they would thereby be privy to information that they would have previously had to wait for. Present practice is to wait with bated breath for the monthly housing indexes to come out.

While this study exquisitely opens the real estate pricing market to Big Data and predictive analytics, there are more avenues of research that researchers must endeavor to do. The researchers of this study did not study the shapes of the variable distributions. They could not attempt to normalize any of the non-parametric distributions to better feed them into some models. They were enamored of SHAP values and eschewed much else. They neither ran pair-wise correlations for the correct distributions, nor did they run a partial correlation. This would be enlightening and would confirm the SHAP values or call into question their validity. Coefficient of determination was never mentioned.

Conclusion

The researchers effectively prove that TOM is the most important feature. Now one must better understand the feature. An SME was incredulous when I described the results of this study to him. He gave two scenarios. One seller priced his property appropriately and the property was sold in short order. Another seller asked for too much and his property did not sell. Eventually this seller had to lower the price to a deflated value because no buyer wanted to buy it even at its fair value. The seller had lost his credibility. Buyers concluded that there must be something wrong with the property because of the magnitude of the TOM. Alternatively, something is wrong with the seller, and this is why the property was mispriced. Nonetheless, buyers are unwilling to pay the 'real' price because of how this seller chose to market it. These outside factors must be considered. Anyone who speaks with a real estate agent will be told that if a house is on the market for more than 60 days, the seller did not ask for the correct price. There are most definitely sellers who are willing to contend with a longer TOM because it pays for them to speculate on making a greater profit. But such models skew our data. We must find data addressing how the sellers priced their properties and why. Then we can better parse the true effect of the TOM variable on price revisions.

References

Casas, P. (2019, March 18). *A gentle introduction to SHAP values in R: R-bloggers*. R. Retrieved April 10, 2023, from https://www.r-bloggers.com/2019/03/a-gentle-introduction-to-shap-values-in-r/

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Grybauskas, A., Pilinkienė, V. & Stundžienė, A. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. *J Big Data* **8**, 105 (2021). https://doi.org/10.1186/s40537-021-00476-0

Lazear, E. P. (1984). *Retail pricing and clearance sales* (No. w1446). National Bureau of Economic Research.

Lundberg, S. M., & Lee, S.-I. (1970, January 1). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems. Retrieved April 10, 2023, from https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Tseng, G. (2018, June 21). *Interpreting complex models with Shap Values*. Medium. Retrieved April 10, 2023, from https://medium.com/@gabrieltseng/interpreting-complex-models-with-shap-values-1c187db6ec83

Yinger, J. (1981). A search model of real estate broker behavior. *The American Economic Review*, *71*(4), 591-605.