Predicción del precio de la vivienda
Comparación de modelos predictivos con CARET

1. Introducción

En este trabajo se presentan modelos predictivos para el precio de un activo de difícil valuación como la vivienda. Se utilizan dos fuentes de datos novedosas para la ciudad de Montevideo: una proveniente de sitios web obtenidos mediante *web scraping* (para el período febrero 2018 - enero 2019) y otra de registros administrativos de transacciones. Se implementan tres modelos fácilmente replicables con el paquete CARET Kuhn (2018): modelo lineal, árbol de regresión y bosques aleatorios, luego se compara su poder predictivo. Los resultados arrojan una mejor performance del modelo de bosques aleatorios (*random forest*) respecto al modelo lineal hedónico, ampliamente difundido en la literatura. Se busca incorporar al análisis de predicción de precios una metodología de creciente aplicación a nivel internacional así como poner a disposición una nueva base de datos procesada y actualizada (ver kaggle dataset).

2. Antecedentes

Los antecedentes se pueden dividir entre los trabajos que tratan sobre el mercado de vivienda en Uruguay y los que aplican técnicas de aprendizaje estadístico para la predicción de precios de vivienda. Para este punto se encontró un solo trabajo nacional y varias referencias internacionales.

Dentro de la segunda categoría, se destaca el trabajo de Mullainathan and Spiess (2017), que realiza una introducción a los modelos de aprendizaje estadístico mediante un ejemplo con datos de precios de vivienda. En éste, ilustran diferentes modelos de predicción y muestran su performance predictiva, destacando la mejor de los modelos de aprendizaje. Este artículo, al igual que Athey (2018) y Varian (2014) fueron la principal motivación para incorporar técnicas de aprendizaje estadístico y resultan referencias básicas para el área económica.

La referencia nacional para este trabajo es Goyeneche et al. (2017), en ella se utiliza una base de datos de tasaciones del Banco Hipotecario del Uruguay con el objetivo de predecir el precio contado de un inmueble. Por otra parte, en el reciente artículo de Čeh et al. (2018) se realiza un trabajo similar al presentado aquí. Se compara la performance predictiva de un modelo de bosques aleatorios en relación a una regresión lineal hedónica para el precio de los apartamentos en Liubliana, Eslovenia.

3. Estrategia metodológica

El objetivo es comparar tres modelos en relación a su poder predictivo. La estrategia de comparación elegida consiste en considerar pocas variables explicativas en ambas bases de datos. La elección de las variables se basa en la confiabilidad de las mismas y en la baja presencia de datos faltantes. La modelización se realiza en cada base de datos por separado y por tipo de propiedad.

Para comparar los modelos, se realiza validación cruzada estándar con el paquete CARET (Kuhn, 2018) y en una muestra de entrenamiento particionada. Basado en los resultados de la validación cruzada, se toman las medidas de performance en la muestra de testeo considerando cada modelo que emerge de la validación cruzada. Las medidas son la menor Raíz del Error Cuadrático Medio (RECM) y el menor Error Porcentual Absoluto Medio (EPAM). Una vez seleccionado el mejor modelo, se incorporan más variables en ese modelo para mejorarlo en términos predictivos (en el caso del desarrollo de bosques se utiliza el paquete randomForest (Liaw and Wiener, 2002)).

4. Datos

Los datos de ofertas fueron recopilados a través de la API (interfaz para acceder a la página web) puesta a disposición por `mercadolibre.com`. Para ello se utilizó un programa elaborado en *python*. Esto sería posible lograrlo a través de R, utilizando paquetes como *rvest* (Wickham, 2019), *httr* (Wickham, 2018), *jsonlite* (Ooms, 2014), entre otros. La muestra original incluye todas las ofertas de venta de inmuebles para la ciudad de Montevideo para el período febrero 2018 - enero de 2019 inclusive. Se realizaron bajadas sucesivas de datos alrededor del día 25 de cada mes.

Luego de la limpieza, la base de datos cuenta con aproximadamente 90.000 observaciones únicas (inmuebles cuyo ID no se repite ni hay indicios de repetición).

La base de datos de transacciones abarca el período enero 2017 - junio de 2018 inclusive e identifica transacciones de compraventa de padrones en Montevideo con destino de vivienda. La muestra seleccionada y procesada cuenta con 12.815 observaciones. A modo de ejemplo, se presenta a continuación la ubicación de las transacciones de apartamentos (las descripciones geográficas utilizan el paquete Leaflet, Cheng et al. (2018) y los gráficos se realizan con ggplot2 Wickham (2016)).

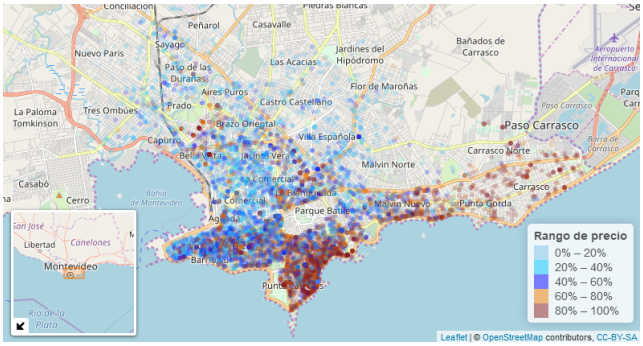

Figura 1: Ubicación de Apartamentos transados

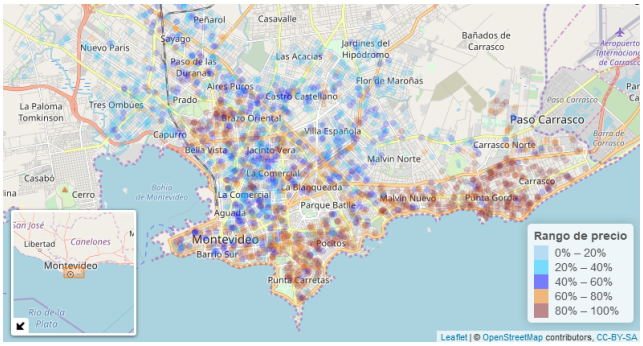

Figura 2: Ubicación de casas transadas

Rango de precios en base a los quintiles por tipo de propiedad.
Fuente: Elaboración propia, en base a DGR, DNC, SIG-IM

5. Resultados

Los resultados arrojan que el modelo de bosques aleatorios tiene una mejor performance predictiva respecto al modelo lineal hedónico, considerando los datos de Montevideo, tanto de ofertas como de transacciones. Esto parecería estar explicado por la naturaleza no lineal del problema de predicción. La superioridad del modelo de bosques aleatorios es aún mayor con los datos de ofertas. Con estos datos para apartamentos, el error porcentual absoluto medio (EPAM) evaluado en una muestra de testeo independiente, se reduce en 40 % (de 17 % a 10 %) al utilizar el modelo de bosques aleatorios respecto al modelo lineal. Para las casas el EPAM se reduce 36 % (22 % a 14 %) respecto al modelo lineal. Las predicciones de los modelos ampliados, en el caso de las ofertas, presentan un error de USD 29.500 (RECM) para apartamentos, y USD 56.800 para las casas. En el caso de transacciones, la performance del modelo de bosques aleatorios también mejora las predicciones respecto al lineal, aunque en proporciones modestas. En el caso de los apartamentos mejora un 6 % (considerando EPAM) y en el caso de las casas la reducción es de 1 % (tomando RECM).

Los resultados obtenidos pueden establecer una base para desarrollos posteriores que exploren variantes en los modelos utilizados, por ejemplo ajustando los parámetros de variables y poda. Igualmente, se podría completar el análisis contemplando nuevas técnicas de aprendizaje estadístico que no se incluyen en el presente trabajo, principalmente Máquinas de Vectores de Soporte. Adicionalmente es posible incorporar más variables hedónicas (incluyendo variables espaciales y geográficas) y utilizar técnicas de econometría espacial. Adicionalmente, los modelos podrían contemplar por un lado la dimensión temporal y por el otro, fundamentos macroeconómicos.

Referencias

Athey, S. (2018). The Impact of Machine Learning on Economics. In *The Economics of Artificial Intelligence: An Agenda*, pages 1–31. University of Chicago Press.

Čeh, M., Kilibarda, M., Lisec, A., and Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, 7(5):168.

Cheng, J., Karambelkar, B., and Xie, Y. (2018). *Interactive Maps with JScript Leaflet*. R package v. 2.0.2.

Goyeneche, J. J., Moreno, L., and Scavino, M. (2017). Predicción del valor de un inmueble mediante técnicas agregativas. *Serie DT IESTA (17/1)*.

Kuhn, M. e. a. (2018). *Caret: Classification and Regression Training*. R package v. 6.0-84.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]*.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2018). *httr: Tools for Working with URLs and HTTP*. R package v. 1.4.0.

Wickham, H. (2019). *rvest: Easily Harvest (Scrape) Web Pages*. R package v. 0.3.3.