clusterAI

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Análisis del mercado inmobiliario con Machine Learning

UNIVERSIDAD TECNOLOGICA NACIONAL - BUENOS AIRES, ARGENTINA

Alvaro Beccaglia
alvarobeccaglia@gmail.com

Lucas Pessagno
pessagnolucas@gmail.com

1 INTRODUCCIÓN Y OBJETIVOS

El mercado inmobiliario es uno de los más importantes para la economía de un país. Los inmuebles son uno de los instrumentos de inversión más extendidos, debido a su baja volatilidad. Comprender cómo funcionan estas tendencias es fundamental para el manejo del capital inversor.

El objetivo del presente trabajo es predecir la compra-venta de inmuebles en la Ciudad de Buenos Aires a partir de la aplicación de los modelos ***Random Forest Regressor***[(1)] y ***Sarima-Arima***[(2)], tomando como ingreso indicadores de acceso público del Gobierno Nacional Argentino y del Gobierno de la Ciudad de Buenos Aires.

2 MATERIALES Y MÉTODOS.

Para poder realizar la predicción en la compra-venta de inmuebles elegimos utilizar el aprendizaje supervisado.

Los modelos utilizados fueron los siguientes:

I. Random Forest Regression.

Los modelos Random Forest están formados por árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generados mediante ***Bootstrapping***[(3)].

En el entrenamiento de un árbol de regresión, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal.

El proceso de entrenamiento de un árbol de predicción (regresión o clasificación) se divide en dos etapas:

i. División sucesiva del espacio de los predictores generando regiones no solapantes (nodos terminales) .
ii. Predicción de la variable respuesta en cada región.

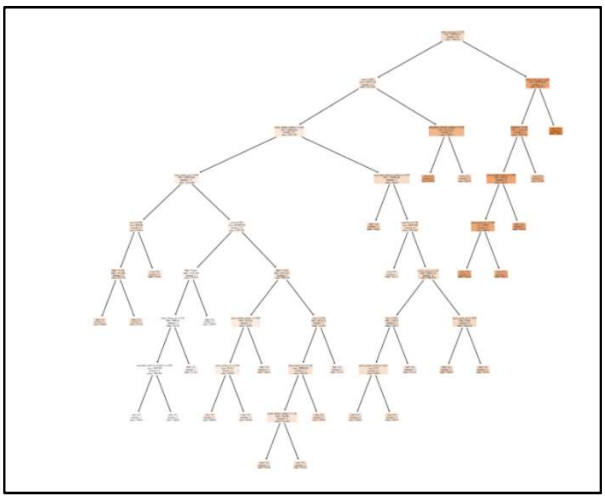

Figura 1. Árbol de decisión modelado por el RFR.

Cuando se quiere predecir una nueva observación, se recorre el árbol acorde al valor de sus predictores hasta alcanzar uno de los nodos terminales. La predicción resultante se obtiene agregando las predicciones de todos los árboles individuales que conforman el modelo.

II. Modelo autorregresivo integrado de media móvil SARIMA – ARIMA.

ARIMA es una clase de modelo que proyecta una serie de tiempo determinada en función de sus propios valores pasados, es decir, sus propios retrasos y los errores de pronóstico previos. Su fórmula de aplicación es la

siguiente:

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Si una serie de tiempo tiene patrones estacionales, entonces necesita agregar términos estacionales y se convierte en SARIMA, abreviatura de "Seasonal ARIMA". En este caso, poseemos un patrón con estas características, por ese motivo se seleccionó un modelo SARIMA.

Este modelo consta de tres partes:

i. ***Parte autorregresiva o "residuos"***: La parte autorregresiva se refiere a la relación entre la variable (que estamos tratando de pronosticar) con sus propios valores rezagados más próximos. Incorpora el "ruido" o "residuo" al pronóstico.

ii. ***Pieza integrada o "estacional"*:** La parte integrada se refiere al orden de diferenciación. Brinda el contexto estacional.

iii. ***Parte de media móvil o "tendencia"*:** Asociado a la media móvil indica la dependencia del valor presente de la variable de la serie temporal de los términos de error rezagados.

Podemos visualizar en la *Figura 2* la descomposición realizada sobre el target (cantidad de compra-venta) donde se diferencian las partes citadas.

Estos patrones estadísticos le permitirán al modelo generar predicciones a futuro.

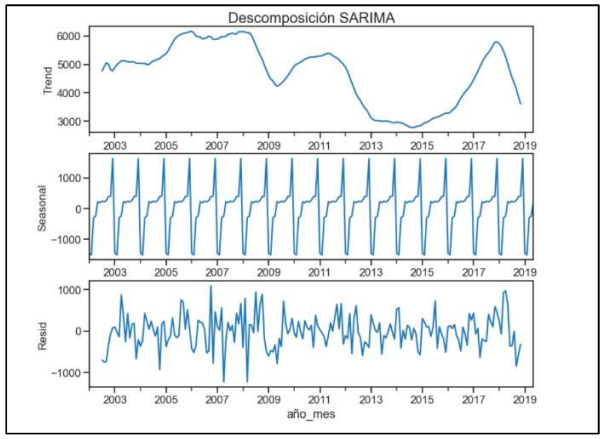

Figura 2. Descomposición SARIMA del Target.

III. Modelo de Regresión polinomial.

La regresión polinomial es un modelo de análisis de regresión en el que la relación entre la variable independiente X y la variable dependiente Y se modela con un polinomio de grado n que posee las siguientes características:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_n X^n + \varepsilon$$

En la *Figura 3* podemos ver la nube de puntos usada para el entrenamiento (dato) y el ajuste del modelo.

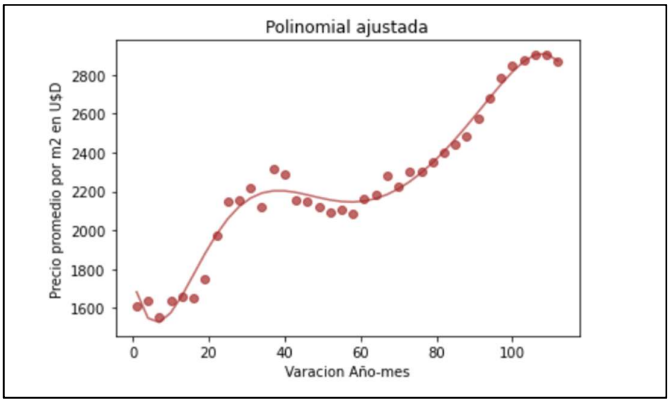

Figura 3. Polinomial ajustada para Precio promedio por m2 en U$D.

3 PREPOCESAMIENTO DE DATOS

La información utilizada para el desarrollo del proyecto, en su mayoría, fue tomada de la página oficial del *Gobierno de la ciudad (Ba Data)*. Las principales variables incluidas en este Dataset fueron:

Precio de venta, Precio de Alquiler, Préstamos hipotecarios UVA y Actas notariales de compra-venta.

Además, para enriquecer el análisis agregamos del sitio oficial del *Gobierno Nacional de Argentina (Datos Argentina): la variación del tipo de cambio, expectativa de inflación y tasas de interés promedio.*

Toda esta información fue preprocesada con el objetivo de unificarla en un dataset para el desarrollo del proyecto de investigación.

4 ANÁLISIS EXPLORATORIO DE DATOS (EDA) E INGENIERÍA DE FEATURES(FE)

En el EDA se plantearon las comparaciones entre distintas **Features** de interés y las compra - venta realizadas **(Target).**

Debido a faltantes en el dataset, comenzamos el estudio acotando el análisis entre Octubre 2010 y Agosto de 2019.

En una primera aproximación con los datos, se puede observar el proceso devaluatorio ocurrido entre Julio y Octubre de 2018. Durante este período, el tipo de cambio pasó de aproximadamente 25$/US$ a más de 35$/US$. Este suceso generó una contracción en el precio de los alquileres en dólares. La tendencia parece indicar que el mercado comenzó a corregir progresivamente la distorsión. (*Figura 4*).

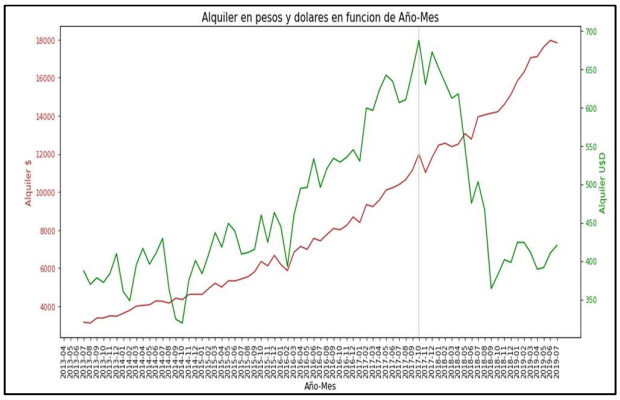

Figura 4. Alquiler en pesos y dólares en función de Año- Mes.

Al analizar la Figura 5 “Alquiler en dólares vs. cantidad total de superficie en alquiler” se puede visualizar una correlación negativa que se mantiene durante casi todo el tramo de estudio hasta el momento de la devaluación mencionada previamente (Octubre de 2018).

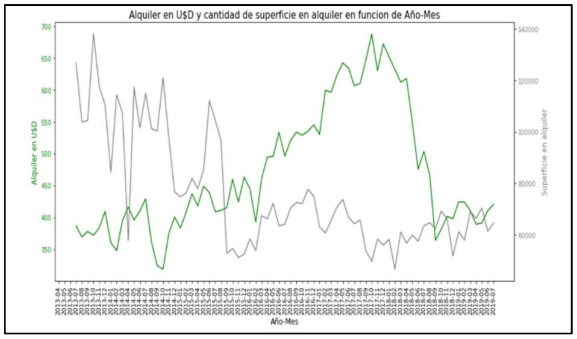

Figura 5. Alquiler en U$D y cantidad de superficie en alquiler en función de Año-Mes.

Luego se analizó la relación entre la variable y la cantidad de compra-venta (Figura 6). Se pudo observar que si bien existe una relación desde el 2012 en adelante (*línea vertical gris en el gráfico*), no se puede decir que es concluyente ya que se puede ver que para valores anteriores no se respeta esta tendencia. El dataset no es lo suficientemente amplio como para afirmar dicha conclusión.

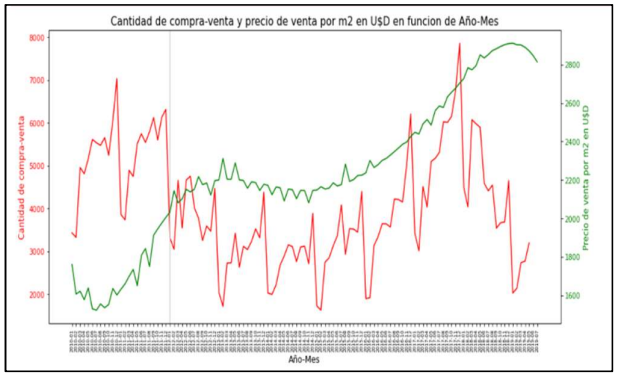

Figura 6. Cantidad de compra-venta y precio de venta por m2 en U$D en función de Año-Mes.

Al relacionar la Cantidad de Ventas y la Expectativa de inflación e interés (Figura 7), se observa que el vínculo es inverso, concluyendo que la cantidad de compra-venta se incrementa en períodos de baja inflación y decrece en períodos de alta inflación.

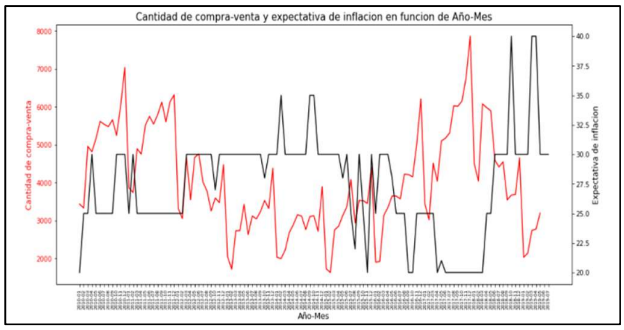

Figura 7. Cantidad de compra-venta y expectativa de inflación en función de Año-Mes.

Al comparar la tasa de interés con el target (Figura 8) se refleja que, al igual que con la inflación, parece existir una relación inversa. Sin embargo, la tasa de interés, a diferencia de la inflación, puede ser controlada con mayor

facilidad por parte del Estado, motivo por el cual lo definimos como una regla “blanda”. Como referencia, durante la gestión nacional del periodo 2015-2019, el gobierno intentó mantener la tasa de interés por encima de la inflación. Esta política deja al descubierto la manipulación de la variable y nos advierte de una posible correlación entre la tasa de interés y la expectativa de inflación.

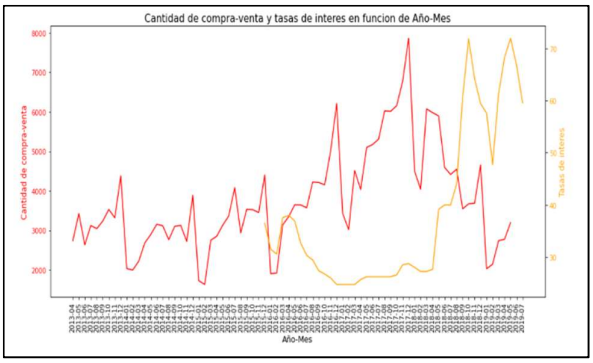

Figura 8. Cantidad de compra-venta y tasas de interés en función de Año-Mes.

La ventana de tiempo utilizada para el proyecto contempla un período previo y posterior al inicio de la política de crédito UVA (Mayo 2016). Motivo por el cual, se decidió analizar su impacto (Figura 9).

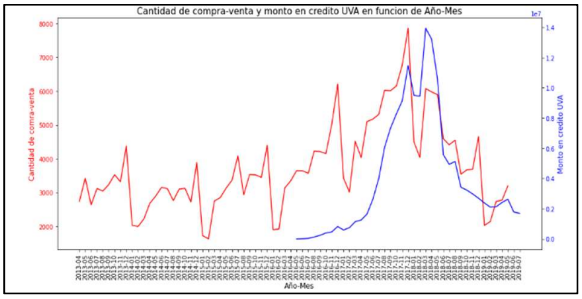

Figura 9. Cantidad de compra-venta y monto en crédito UVA en función de Año-Mes.

Se puede observar cómo la política rompió la tendencia de compra-venta con una gran alza hasta su pico máximo de 8000. Sin embargo, los valores previos eran comparativamente bajos con respecto al resto del periodo analizado. Esto llevó a comparar mediante la técnica de boxplot las distribuciones de compra-venta entre ambos periodos.

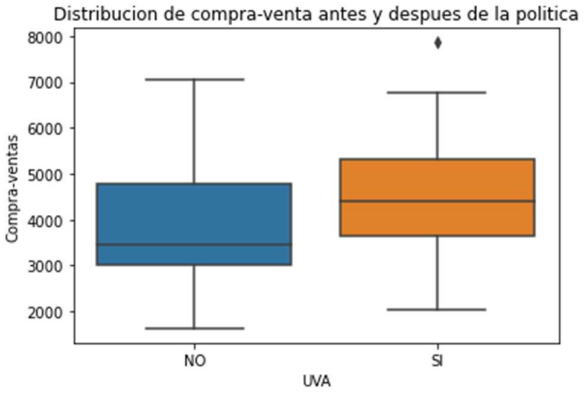

Figura 10. Distribución de compra-venta antes y después de la política.

Se puede visualizar que, si bien existió una suba estacional y marcada, las distribuciones no difieren demasiado, ubicando sus medianas en 3818 ventas mientras no existía la política y 4501 en su vigencia.

Durante el desarrollo del EDA y para la utilización de los datos en el entrenamiento de los modelos se han realizado las siguientes acciones de FE:

- **i.** Reemplazo de Nulls por valores obtenidos del ajuste del modelo de regresión polinomial en la feature requerida.
- **ii.** Reemplazo de Nulls por 0.
- **iii.** Reemplazo de Nulls por datos obtenidos en otras bibliografías.
- **iv.** Agregado de una columna que permite la vinculación del dataset con la estacionalidad.
- **v.** Acotación del dataset.Split y escalamiento de features

5 EXPERIMENTOS Y RESULTADOS

Una vez procesado el dataset, se probaron tres modelos de regresión con sus correspondientes GridSearch (búsqueda de grilla) con el propósito de obtener los mejores hiperparamétros y atenuar el error de predicción. Los resultados fueron los siguientes:

```
Mean Absolute Error for model:
Random Forest Regressor:396.69
Bayesian Regression
Ridge:440.03
Support vector machine:1149.74
```

Por este motivo se seleccionó el Random Forest Regressor. Luego, se re-entrenó el modelo con los hiperparametros obtenidos alcanzando los siguientes errores:

```
Random Forest Regressor:
Mean Absolute Error:396.69
Mean Squared Error:157365.78
R-cuadrado:0.8857
```

En la *Figura 11*, se pueden evaluar las diferencias entre los valores predichos y los valores reales de entrenamiento:

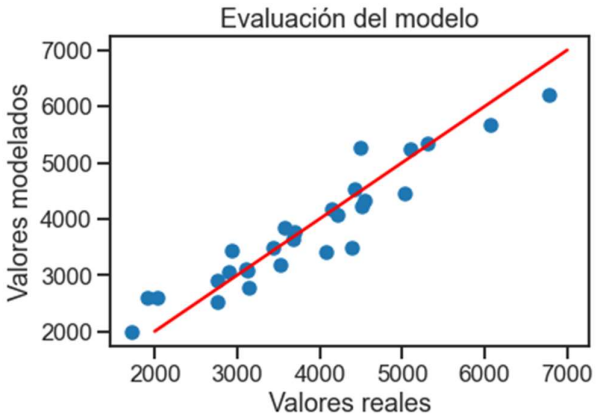

Figura 11. Evaluación del modelo de Random Forest Regressor.

Para proyectar valores de compra-venta en función de valores históricos se utilizó un modelo SARIMA. Obteniendo la proyección que se puede visualizar en la *Figura 12*.

Según esta, para diciembre del 2020 se realizarán alrededor 2500 compra-ventas. Cabe destacar que este modelo resulta útil para plazos acotados de análisis.

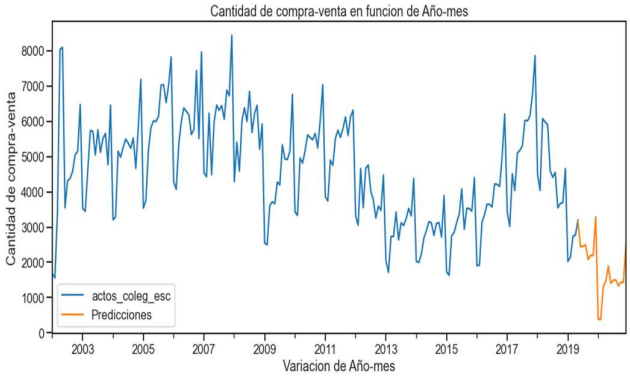

Figura 12. Proyección del modelo de SARIMA.

6 CONCLUSION

A partir de los datos provistos se ha logrado desarrollar la FE requerida y el entrenamiento de los modelos que permiten predecir, con un cierto margen de error, **la cantidad de compra-ventas en la Ciudad de Buenos Aires en función de indicadores de carácter público para un momento o plazo determinado**.

Algunos de los indicadores utilizados dependen de manera directa de las políticas de Estado, lo cual, puede producir perturbaciones en proyecciones futuras. Con el fin de robustecer el modelo, sería recomendable su actualización periódica y la incorporación de nuevas variables más independientes.

La utilización del modelo **SARIMA** permite **proyectar, a partir de la tendencia histórica, posibles valores de compra-venta que luego pueden ser corroborados con el modelo de regresión lineal**. De esta manera, conocer (o proyectar) los indicadores económicos del período a analizar permitirá validar la predicción del primer modelo.

Este análisis podría resultar útil tanto para el sector público, ya sea midiendo el alcance e impacto de una política, como para el sector privado, por ejemplo, si una cadena de inmobiliarias desea estimar sus ingresos por compra-venta en un plazo definido.

References:

(1) Mark R. Segal - ***Machine Learning Benchmarks and Random Forest Regression***

(2) Prof. Rafael de Arce. Prof. Ramon Mahia (1970)- ***Modelos Arima***

(3) D. A. Freedman (1981) – ***Bootstrapping Regression Models***