Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Metodos Numericos 1C 2021

Trabajo Práctico 3: Metros Cuadrados Minimos Lineales

Grupo 16

Integrante	LU	Correo electrónico
Cantini Budden, Sebastian	576/19	sebascantini@gmail.com
Kruger, Lucas Ivan	799/19	lucaskruger10@gmail.com

Resumen:

En el presente informe buscamos estudiar como diferentes variables afectan el precio de inmuebles ubicados en México. Utilizamos Segmentación para comparar precios y otras variables que definen las viviendas, sea en todo el país o en una ciudad en particular. Tomaremos a consideración aspectos como metros cuadrados, palabras positivas en la descripción y ubicaciones de dichos inmuebles para ver como predecir el precio. Utilizaremos el error cuadrático medio para analizar la correctitud de nuestras aproximaciones.

Palabras Clave

Cuadrados Minimos Lineales,Error Cuadrático Medio, Inmuebles, Ecuaciones Normales, Segmentación

I. Introducción Teórica

En el presente informe vamos a estudiar algoritmos de predicción de características de inmuebles. Utilizaremos un algoritmo de clasificación supervisado, el cual sera entrenado con una base de avisos de ventas de inmuebles con precios conocidos, esto servirá para aproximar unas características a partir de otras. Es decir, se puede fijar una característica y tratar de explicarla mediante otras. Usamos un conjunto de datos de avisos inmobiliarios de México.

A. *Cuadrados Mínimos Lineales:*

Para aproximar las características utilizaremos la técnica de Cuadrados Mínimos Lineales (CML).
Dada una familia de funciones $\{\phi_1, \phi_2, ..., \phi_n\}$ y un conjunto de vectores $\{x_1, x_2, ..., x_m\} \subseteq R^k$ y escalares $\{y_1, y_2, ..., y_m\} \in R$ entonces tomamos $M \in R^{mn}$ como la matriz con elementos $m_{ij} = \phi_j(x_i)$ e $y \in R^m$ como el vector compuesto por los escalares y_i. A cada componente de un vector x lo llamamos feature y cada vector representa una medición de estos features. Es por eso que a cada medición x_i le corresponde un resultado y_i. Entonces, cuadrados mínimos es el problema que consiste en hallar el vector α que minimice $\|M\alpha - y\|$. Lo que esto logra es hallar la mejor combinación lineal de funciones ϕ_i que aproximan los valores de y con los vectores x asociados. En particular, para regresión lineal consideramos, $\phi_1 = 1$ y $\phi_i = e_{i-1}^t \; \forall i \neq 1 \leq k+1$, donde e_i es el i-esimo vector canónico. En otras palabras, para regresión lineal consideramos una combinación lineal de los features de las mediciones. Una manera de resolver el problema de cuadrados mínimos es usando ecuaciones normales. Resulta que:

$$\alpha = min_\beta(||M\beta - y||_2) \iff M^t M\alpha = M^t y \quad (1)$$

Nos basaremos en la resolución de este sistema para nuestro algoritmo futuro.

Algo a tomar en cuenta es que utilizar esta técnica con sets muy grandes, perdemos especificidad en los resultados, por lo que utilizaremos el método de segmentación como veremos mas abajo.

Utilizaremos las siguientes métricas para ver que tanto difieren los precios estimados de los reales:

B. *RMSE:*

Root Mean Square Error; el error cuadrático medio (ECM o RMSE en ingles) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. Esta métrica pesa mas las muestras con valores altos que las muestras con valores bajos. Dado un modelo y una observación $(x_{(i)}, y_{(i)}$, definimos $\hat{y}_{(i)}$ y $e_(i) = y_{(i)} - \hat{y}_{(i)}$

$$RMSE(\hat{f}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_{(i)}^2} \quad (2)$$

C. *RMSLE:*

Root Mean Square Logarithmic Error (RMSLE) utiliza el mismo concepto que RMSE, en el sentido de que intenta estimar error de la misma forma. RMSLE intenta corregir una desventaja masiva que tiene el RMSE, ya que en este ultimo no es ideal para toda situación. La métrica RMSE pesa mas las muestras con valores altos que las muestras con valores bajos. Esto significa que si intentamos mejorar un 10% el error de nuestra muestra, este cambio, tiene mucha mas influencia en un conjunto de datos con valores altos que una muestra con valores bajos. Lo que RMSLE intenta lograr es usar logaritmos para que el crecimiento de los datos no sea lineal y así disminuir la debilidad que presenta RMSE. RMSLE utiliza la siguiente ecuación para calcular el error:

$$RMSLE(\hat{f}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(log(y_{(i)}+1) - log(\hat{y}_{(i)}+1))^2} \quad (3)$$

D. *Segmentacion:*

Debido a que es muy complejo tratar de explicar un dataset tan grande con una única función, vamos a segmentar el dataset y luego aplicar CML(cuadrados mínimos lineales). El set se puede segmentar de varias maneras (ej: categorías, lugar, etc,) y de esta forma producir varias aproximaciones para tratar de explicar los datos.

E. *Feature Engineering:*

El proceso de feature engineering consiste en producir nuevas características para utilizar en los métodos de aproximación. La idea principal es concentrarnos en particularidades en los datos para ver correlaciones entre estos y las variables numéricas en nuestra experimentación, o para utilizar para la segmentación para ver si se magnifican nuestros resultados.

F. *Variables:*

En el trabajo practico utilizaremos dos tipos de variables distintas:

1) Variables Numéricas: Estas variables son las que se pueden cuantificar, sea cantidad de baños, cantidad de metros cuadrados o precio. Cualquier aspecto al cual se le puede dar un valor numérico.

2) Variables Características: Estas otras variables son las que no se pueden cuantificar, sea nombre de la calle, ciudad o la descripción de la propiedad. Es importante hacer la siguiente observación. Es posible que una variable característica tenga un valor numérico, sin embargo, no hay noción de distancia entre estas, como por ejemplo el ID de la propiedad, no indica ningún tipo de significado mas que una forma de identificar el inmueble. Adicionalmente algunas variables características se pueden representar con números, como por ejemplo, el tipo de inmueble, con una casa tomando el valor uno, un departamento el valor dos, etc, todo con el fin de distinguir, pero en este caso, tampoco es considerado como una variable numérica.

Nuestros conjuntos de datos incluyen las siguientes variables:

- ID: El identificador del inmueble es una variable característica que se utiliza para diferenciar una propiedad de otra. Esta variable es importante ya que hay casos, como en un complejo de departamentos, donde dos inmuebles coinciden en todas las otras variables.
- Información de la publicación: Un conjunto de variables que nos dice el titulo, una breve descripción del inmueble con algunas palabras claves y fecha de publicación.
- Tipo de Propiedad: Describe si el inmueble es una casa, apartamento, terreno, etc.
- Ubicación: Variables características como provincia, ciudad y dirección del inmueble como también latitud y longitud.
- Variables Numéricas: Nos encontramos con cantidad de baños, habitaciones, metros cubiertos y totales, piscinas, garajes y precio del inmueble.
- Cercanías: Estas variables nos hablan sobre cuantas de las siguientes estructuras tiene cerca: escuelas, centros comerciales, etc.

G. *Outliers:*

En este informe hablaremos mucho de outliers, por ende, queremos dar una breve definición de lo que son para nosotros en el contexto del informe y experimentos. Outliers son aquellos datos mas extremos en una muestra, que puede afectar potencialmente a la estimación de los parámetros del mismo. Para detectarlos quitaremos aquellos que se alejen 2 veces el desvió estándar de la predicción. Lo que haremos sera Nosotros quitaremos outliers con Python y veremos mas adelante, como quitar estos afecta el RMSE y RMSLE.

II. Hipótesis y Datos

En esta sección propondremos diferentes hipótesis y experimentos para estudiar nuestros datos.

A. *Hipótesis:*

Primera Hipótesis: Metros Cuadrados y Precio

Nuestra primera hipótesis habla sobre como el tamaño de la propiedad afecta al precio. Creemos que cuanto mas grande sea la propiedad mas valor tendrá. Notar que cuando hablamos de tamaños, hablamos de metros totales, ya que si solo consideramos los metros cubiertos nos podemos cruzar con algunas viviendas que son mas chicas, sin embargo, estas pueden tener patio o terraza aumentando el precio del inmueble. Por esa razon, ademas proponemos que al aumentar la suma de los metros totales con los metros cubiertos conseguiremos una prediccion aun mas ajustada.

Segunda Hipótesis: Habitaciones, baños, garaje y Metros Cubiertos

Pensamos que las habitaciones, baños y garajes sirven para predecir la cantidad de metros cubiertos de un inmueble. Ya que estas tres variables numéricas representan los posibles espacios cubiertos creemos que una combinación de estas sera una gran forma de predecir la cantidad de metros cubiertos de tal propiedad. Por ende nuestra hipótesis es que al aumentar el valor de la suma de estos tres valores, los metros cubiertos aumentaran.

Tercera Hipótesis: Cercanía a la Playa y Precio

Nuestra tercer hipótesis habla sobre una particularidad geográfica. No es secreto que a muchas personas les gusta vacacionar cerca de la costa. Por eso planteamos que el precio subirá mientras la propiedad se acerque mas a la costa. Creemos que la comodidad adicional que brinda acortar la distancia es justificativo suficiente para un aumento de precio.

B. *Datos:*

El conjunto de datos son los provistos por la cátedra. En estos nos encontramos con 240000 viviendas que tiene todas las variables numéricas y categóricas vistas anteriormente. A algunos de los inmuebles les falta variables. Para resolver este problema, en Python al segmentar, ignoramos las filas con casillas vacías de las variables que utilizaremos. Para todos los experimentos tomaremos una décima parte del conjunto de entrenamiento para testeo.

C. *Implementación:*

Implementamos nuestro código de cuadrados mínimos lineales en C++ y analizamos nuestros datos usando diversas librerías de Python en nuestra notebook[1]. Esta implementación siguió la idea general planteada en la sección Cuadrados Minimos Lineales. Utilizaremos segmentación para resolver el problema planteado anteriormente en la que CML intenta explicar todos los datos con una única función. La segmentación se encargara de generar conjuntos de datos más homogéneos y controlados para que cuando hagamos experimentos, los errores calculados sean menores y más fáciles de interpretar según el experimento.

III. Resultados y Discusión

A. *El Precio y los Metros:*

Como dijimos anteriormente creemos que la relación del tamaño con el valor del inmueble es lineal, ya que se espera un crecimiento notorio de precio mientras que aumenta la cantidad de metros cuadrados. Experimentaremos segmentando en "Casas".

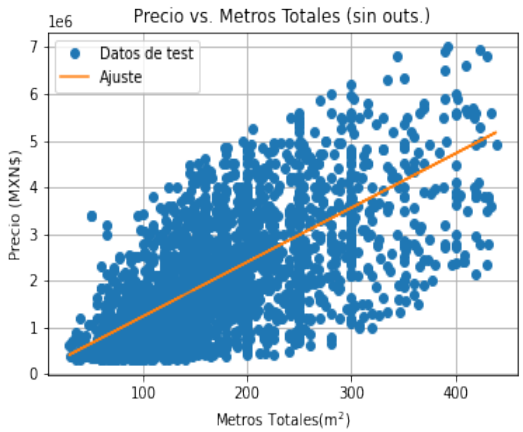

Fig. 1. Experimento donde se mide la relación entre precio y metros cuadrados totales, sin outliers. Realizado sobre un test con 32816 datos, de los que usamos 3281 para test. Estos datos se consiguen segmentando en casas y haciendo dropna en el conjunto total de datos.

Para nuestra sorpresa, nos encontramos con mas variación de datos de los que creíamos. En la imagen fig.1 podemos ver que cuantos mas metros cuadrados hay, mas inconsistencia hay en los datos. En otras palabras hay mas variación cuantos mas metros cuadrados hay.

Hay muchas razones de por que esto puede ser así. La primera razón, y capaz la mas obvia, es que nosotros estamos relacionando una variable con el precio. Anteriormente vimos que las publicaciones de los inmuebles aportan muchos datos distintos para que la persona interesada sepa lo mas posible del lugar. Sin duda estas variables adicionales que no estamos tomando en cuenta afectan el precio, como por ejemplo cantidad de habitaciones, los metros cubiertos, o si tiene o no cochera. Es por esto que dos inmuebles del mismo tamaño puedan tener precios totalmente distintos.

Otra razón es que, como todo lugar, hay partes mas caras que otras. Entonces algo como por ejemplo ciudad también puede afectar el precio del inmueble. Usaremos segmentación para analizar los datos en las ciudades de Zapopan (fig.2) y Querétaro (fig.3) para intentar de quitar esta variable y poder ver si existe un patrón mas lineal y menos disperso. Elegimos estas ciudades en particular ya que son las que mas datos nos brindan de nuestro conjunto de datos. Por lo visto en la notebook [1] Sabemos que Zapopan tiene una mayor cantidad de casas pequeñas (casas entre $25m^2$ a $430m^2$) Querétaro (casas entre $50m^2$ a $850m^2$) tiene una concentración en casas mas grandes. Esto nos permitirá ver dos conjuntos de datos distintos para ver si los errores y optimizaciones planteadas funcionan para ambas distribuciones.

Como se puede ver ambos gráficos presentan el mismo patrón, un crecimiento lineal. Se puede ver claramente que Zapopan tiene mas dispersión que Querétaro, similar a fig.1. Esto puede ser por la razón dada anteriormente. Querétaro, por el otro lado, tiene mas similitudes con lo que nosotros creíamos que pasaría; al limitar los puntos a una ciudad particular, los datos serian mas consistentes. Ambos, tienen menos dispersión que el que tiene todos los conjuntos de datos. Esto se debe a que, como dijimos antes, hay ciudades mas caras que otras. Por ende, segmentar en ciudades nos ayudo a ver una mejor predicción de precios, por que nos permitió utilizar un conjunto de datos mas homogéneo.

Los puntos de ambas ciudades, en conjunto, forman un cono, con la punta en la parte inferior izquierda del gráfico. Esto nos indica que el precio de los inmuebles pequeños es mas consistente. Lo que implica que la dispersión principal de datos ocurre en las viviendas mas caras.

Si vemos bien los gráficos de cada ciudad, podemos ver que Zapopan tiene precios mas altos que Querétaro. Y como dijimos, cuanto mas alto es el precio, mas dispersión de datos vamos a tener. Similar al problema con RMSE, cuantos mas grandes son los números, mas error hay. Aquí el error es la diferencia de precios entre las viviendas. Esto es así para discriminar los beneficios

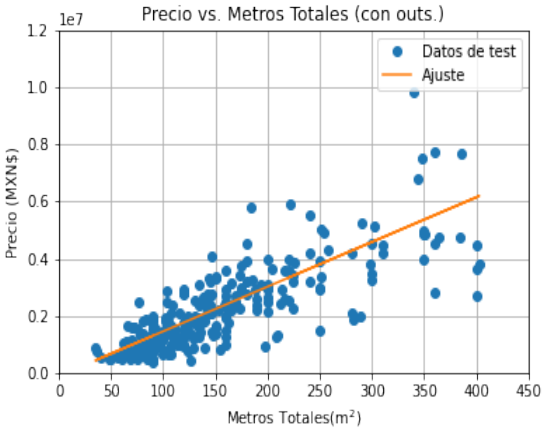

Fig. 2. Experimento donde se mide la relación entre precio y metros cuadrados totales en Zapopan (Casas), con outliers.

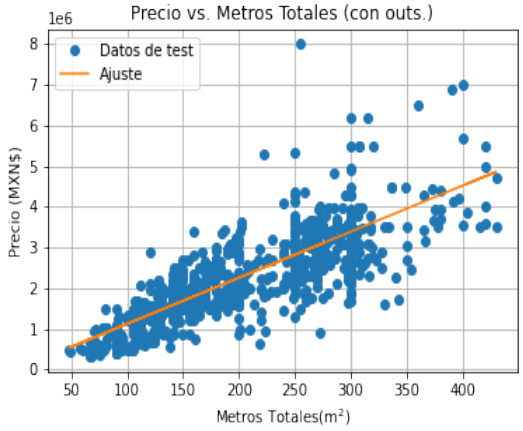

Fig. 3. Experimento donde se mide la relación entre precio y metros cuadrados totales en Querétaro (Casas), con outliers.

de un inmueble y de otro, ya que un cambio significante de precio para un inmueble chico (con precio bajo) pasa a ser insignificante para los precios mayores que ofrecen las propiedades mas grandes.

*1) **Análisis del Ajuste**:* Para verificar la precisión del ajuste, utilizando las ecuaciones (1) y (2), pudimos calcular el RMSE y el RMSLE respectivamente de cada uno de los gráficos superiores. Calculamos las métricas de cuatro formas distintas. Primero vemos las métricas de los datos sin ningún cambio. Luego vimos las métricas utilizando el método de K-Fold Cross Validation, con $K = 10$, en el conjunto de datos de entrenamiento. El tercer método fue sin hacer K-Fold Cross Validation pero sacando los outliers. Los outliers generan error adicional, sin embargo por razones de consistencia optamos por sacarlos para ver el caso mas general. El ultimo método (y el que mostraremos mas adelante) es una combinación del método dos y el método tres. Utilizamos K-Fold Cross Validation y adicionalmente sacamos los outliers. Creemos que estas son las mejores métricas para utilizar. Primero veamos las métricas sin sacar outliers para poder comparar:

- General:
  - $RMSE : (1.279 \pm 0.045) \times 10^6 MXN\$$
  - $RMSLE : (0.602 \pm 0.009) MXN\$$
- Zapopan:
  - $RMSE : (9.297 \pm 1.311) \times 10^5 MXN\$$
  - $RMSLE : (0.404 \pm 0.032)$
- Querétaro:
  - $RMSE : (7.443 \pm 0.433) \times 10^5 MXN\$$
  - $RMSLE : (0.338 \pm 0.013)$

Los resultados que obtuvimos sacando los outliers son los siguientes:

- General:
  - $RMSE : (9.101 \pm 0.181) \times 10^5 MXN\$$
  - $RMSLE : (0.542 \pm 0.009)$
- Zapopan:
  - $RMSE : (6.775 \pm 0.375) \times 10^5 MXN\$$
  - $RMSLE : (0.367 \pm 0.028)$
- Querétaro:
  - $RMSE : (5.532 \pm 0.147) \times 10^5 MXN\$$
  - $RMSLE : (0.305 \pm 0.012)$

Recordemos que las métricas hablan sobre la predicción de los precios. Como los precios andan en los millones de pesos, y estos valores altos son la debilidad de RSME como vimos en la seccion superior, Estos números altos generan que la varianza sea mucho superior y como este aspecto no afecta demasiado a RMSLE, estos valores y su varianza son mucho mas pequeños.

Comparando ambos sets de resultados podemos ver que quitando los outliers, nuestras métricas obtienen una mejora notoria.

*2) **Aumento de Variables**:* Siguiendo la idea anterior (respecto a los conos que se forman), veamos si es posible, utilizando una variable adicional como metros cubiertos, reducir la forma de cono que tienen nuestros gráficos. Analizaremos RMSE y RMSLE quitando outliers:

- Zapopan (m^2 Totales):
  - $RMSE : (9.1 \pm 0.665) \times 10^5 MXN\$$
  - $RMSLE : (0.337 \pm 0.013)$
- Querétaro (m^2 Totales):
  - $RMSE : (5.274 \pm 0.426) \times 10^5 MXN\$$

  - $RMSLE$: (0.235 ± 0.016)
- Zapopan (m^2 Totales + m^2 Cubiertos):
  - $RMSE$: $(6.595 \pm 0.163) \times 10^5 MXN\$$
  - $RMSLE$: (0.302 ± 0.012)
- Querétaro (m^2 Totales + m^2 Cubiertos):
  - $RMSE$: $(3.836 \pm 0.106) \times 10^5 MXN\$$
  - $RMSLE$: (0.206 ± 0.01)

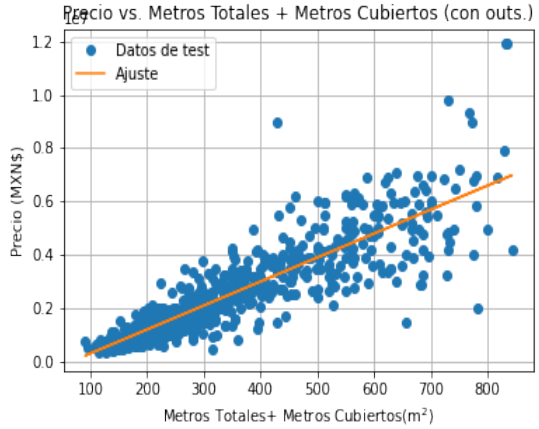

Fig. 4. Experimento donde se mide la relación entre precio y la suma de metros cuadrados totales y cubiertos en Zapopan(Casas),con outliers.

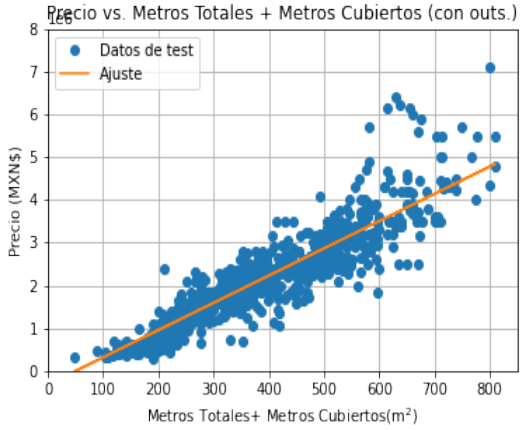

Fig. 5. Experimento donde se mide la relación entre precio y la suma de metros cuadrados totales y cubiertos en Querétaro(Casas), con outliers.

Como podemos ver (fig.4 y fig.5) hay una mejora significante en Querétaro y una mas leve en Zapopan al predecir el precio a base de la suma de estas dos variables. Viendo los gráficos anteriores también podemos ver como el ángulo del cono se redujo bastante logrando menos varianza y también menos error, tal como podemos ver en los RMSE y RMSLE anteriores. Esto se debe a que ahora consideramos aquellas superficies cubiertas que también aumentan el precio.

B. Habitaciones, baños, garaje y Metros Cubiertos

Utilizaremos las ciudades anteriores, Querétaro *(8150 datos de train y 906 datos de test)* y Zapopan *(6891 datos de train y 768 de test)* para analizar si la cantidad de habitaciones, baños y si tiene o no garaje, sirven para intentar de predecir la cantidad de metros cuadrados cubiertos de un inmueble.

Lo que decidimos hacer para unir estas variables es hacer una suma de las cantidades y ver como, de esta forma, logramos predecir adecuadamente el tamaño de la propiedad.

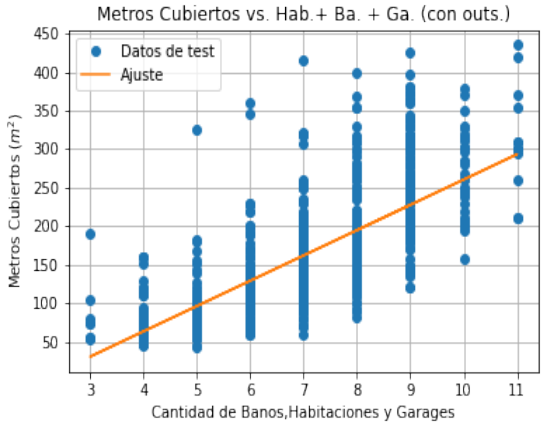

Fig. 6. Relación de la suma de las métricas y los metros cubiertos en Zapopan

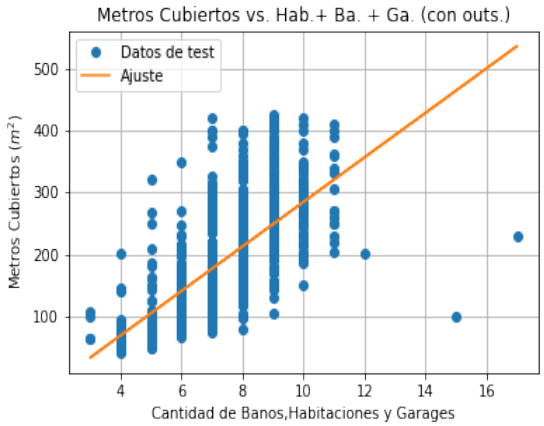

Fig. 7. Relación de la suma de las métricas y los metros cubiertos en Querétaro

Con Outliers:

- Zapopan:
  - $RMSE$: $(57.879 \pm 1.907) m^2\$$
  - $RMSLE$: (0.370 ± 0.02)
- Querétaro:
  - $RMSE$: $(59.8 \pm 1.44) m^2\$$

- $RMSLE$: (0.32 ± 0.013)

Sin Outliers:

- Zapopan:
  - $RMSE$: $(45.538 \pm 0.959) m^2\$$
  - $RMSLE$: (0.31 ± 0.017)
- Querétaro:
  - $RMSE$: $(48.98 \pm 1.129) m^2\$$
  - $RMSLE$: (0.295 ± 0.006)

Como podemos ver en (fig.6 y fig.7) parece haber un aumento lineal en los datos, aunque con una varianza grande. Esto puede deberse a que si bien es lógico esperar el aumento de metros con el de esta suma de variables, puede que una casa tenga habitaciones pequeñas o grandes. Es importante notar que es mejor comparar los resultados usando RMSE ya que los valores son muy pequeños. Estos datos apoyan nuestra hipótesis ya que es cierto que a medida que la cantidad de habitaciones de un inmueble hay un aumento de tamaños notorio, por lo que es posible predecir los metros cubiertos con estas variables.

C. *Cercanía a la Playa:*

Para nuestro ultimo experimento, queríamos evaluar el incremento de precio en relación a la aproximación a la playa. Para esto segmentamos para quedarnos con los datos de Cancún *(1220 casos de train y 136 casos de test)* y el conjunto de Altamira, Ciudad Madero, Miramar y Tampico *(606 casos de train y 68 casos de test)*. Para poder testear esta hipótesis decidimos elegir ciudades sobre el mar que tengan abundante cantidad de datos para poder comparar y verificar datos. Luego generamos una nueva variable numérica; cercanía a la playa. Para crear esta variable, utilizamos la longitud y latitud, poniéndole cotas superiores e inferiores para contener las ciudades, dentro de estos cuadrados seleccionado compararemos la longitud y longitud de cada inmueble con una linea arbitraria a la cual llamaremos *linea de playa*. Utilizamos Google Maps [3] para obtener las cotas que rodean a Cancún y para elegir la linea de playa, tal como vemos en fig.8. En el caso de fig.9 la linea de playa considerada es la linea amarilla del mapa.

Por lo visto en fig.10 y fig.11 podemos decir que nuestra hipótesis no es del todo correcta. En el gráfico de Cancún si se puede ver un crecimiento como creíamos, sin embargo la mayoría de los datos (y la linea de ajuste) muestran que el incremento es mínimo. La mayoría de los datos no muestran índices de cambio de precios.

El gráfico de las otras ciudades muestran algo similar. Este si muestra índices de aumento pero no tanto como esperábamos. Y hay puntos con valores altos en todo el gráfico sin importar distancia. No se puede comprender ninguna forma particular de fig.11 que nos ayude a interpretar la relación que buscábamos.

En Resumen, podemos decir que nuestra hipótesis es parcialmente cierta. Es correcto que hay valores de viviendas que aumentan mientras nos acercamos a la orilla, sin embargo, es falso que este incremento es lineal, ya que la mayoría de los datos no muestran signos de cambios.

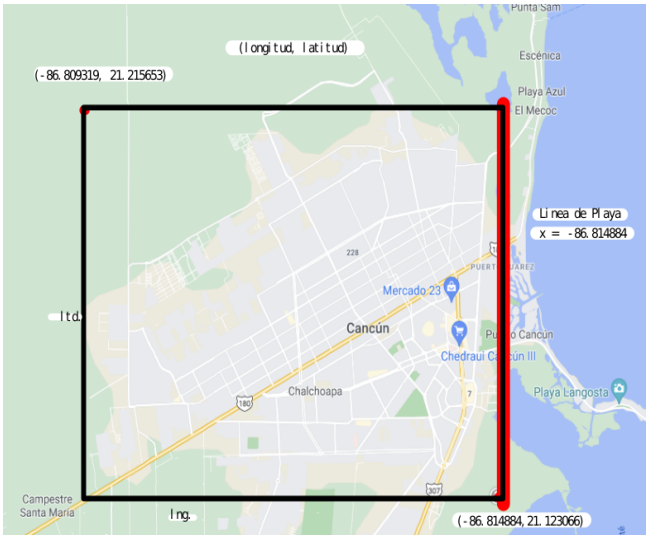

Fig. 8. Mapa de Google Maps con cotas y linea de playa de Cancún.

Fig. 9. Mapa de Google Maps con cotas y linea de playa de Altamira, Ciudad Madero, Miramar y Tampico.

A continuación veremos los RMSE y RMSLE de los datos analizados anteriormente: **Con Outliers:**

- Cancún:
  - $RMSE$: $(1.448 \pm 0.152) \times 10^6 MXN\$$

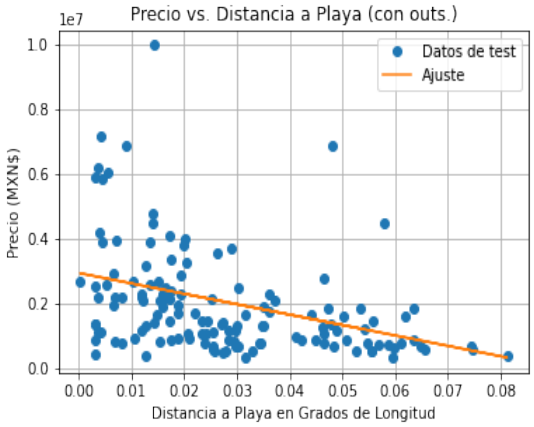

Fig. 10. Experimento donde se mide la relación entre precio y distancia a la playa de Cancún, con outliers.

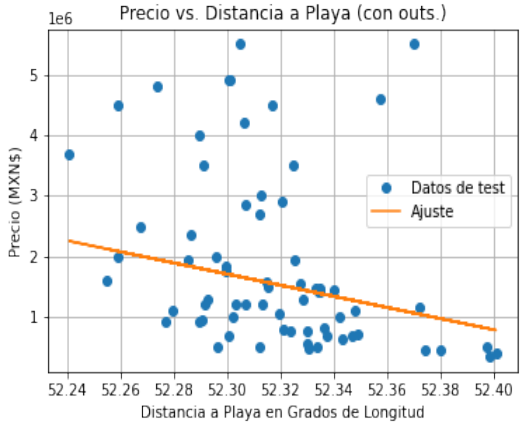

Fig. 11. Experimento donde se mide la relación entre precio y distancia a la playa de Altamira, Ciudad Madero, Miramar y Tampico, con outliers.

  - $RMSLE$: (0.686 ± 0.055)

- Altamira, Ciudad Madero, Miramar y Tampico:
  - $RMSE$: $(1.205 \pm 0.19) \times 10^6 MXN\$$
  - $RMSLE$: (0.725 ± 0.044)

Sin Outliers:

- Cancún:
  - $RMSE$: $(1.021 \pm 0.065) \times 10^6 MXN\$$
  - $RMSLE$: (0.608 ± 0.051)
- Altamira, Ciudad Madero, Miramar y Tampico:
  - $RMSE$: $(8.511 \pm 0.654) \times 10^5 MXN\$$
  - $RMSLE$: (0.646 ± 0.033)

Como se puede ver, en Cancún hay bastante error, esto se debe por los datos mas cercanos a la playa. Como dijimos, hay muchas propiedades que si tienen un gran incremento de precio, sin embargo, la mayoría no. La diferencia entre los comportamientos vistos genera gran dificultad para que el ajuste pueda predecir adecuadamente.

Por otro lado, en el segundo gráfico, tenemos otros problemas que nos generan error en la predicción. Principalmente la falta de forma y la distribución visualmente aleatoria hacen imposible identificar un patrón para el cual un ajuste tendría sentido. Todo esto puede deberse a que estos fenómenos se den mas en ciudades muy turísticas.

IV. Conclusiones:

Analizaremos las conclusiones respecto a cada hipótesis:

- **Primera Hipótesis: Metros Cuadrados y Precio.** Planteamos que ”cuanto mas grande sea la propiedad mas valor tendrá”. En nuestras observaciones pudimos corroborar que es cierto lo que planteamos. Pudimos ver en dos casos particulares que el crecimiento parecía ser lineal. Además propusimos que habría una mejor estimación si utilizábamos la suma de metros totales y cubiertos para predecir el precio, cosa que también parece ser cierta, aunque en casos la mejora es muy pequeña.
- **Segunda Hipótesis: Habitaciones, baños, garaje y Metros Cubiertos.** Planteamos que ”..al aumentar el valor de la suma de estos tres valores(habitaciones+baños+garajes), los metros cubiertos aumentaran.” y los experimentos apoyaron esta hipótesis. Pudimos ver como si había un aumento relacionado, aunque con cierta varianza. Esto se debe a que los metros cubiertos son, por lo general, la suma de estas partes. La varianza puede deberse a la posible diferencia de tamaños entre las habitaciones (refiriéndome a los tres).
- **Tercera Hipótesis: Cercanía a la Playa y Precio.** Planteamos que ” el precio subirá mientras la propiedad se acerque mas a la costa”. Luego del experimento nos dimos cuenta que nuestra hipótesis es parcialmente correcta. Si hay un crecimiento de precios en algunos casos (en la ciudad mas turística) pero la mayoría de los datos implican que la cercanía de la playa es indistinta en el precio. Por ende cercanía no es un buen modelo para el cual intentar predecir el precio de un inmueble.

Concluimos que usar el método de cuadrados mínimos lineales en la estimación de diferentes características en anuncios inmobiliarios es una herramienta con resultados coherentes y acompañada con una buena segmentación, es posible obtener un conjunto de datos mas controlado y homogéneo, que logra generar un ajuste con error

muy bajo. Lo que implica que es una herramienta muy poderosa para análisis de datos. Quitar los outliers redujo el error hasta un 30%. Lo que muestra la importancia de tener estos en consideración al aproximar los datos.

REFERENCES

[1] Notebook incluida en el git.
[2] Infobae: Ciudades Mas Caras
[3] Google Maps
[4] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," Technometrics, vol. 16, 1974.
[5] Francisco Javier Marco Sanjuán (07 de noviembre, 2018). Outlier. Economipedia.com