Análisis, visualización y predicción de Precios Airbnb

Martín Navarrete Villegas
Facultad De Ciencias Fisicomatemáticas
Universidad Autónoma de Nuevo León
Correo electrónico: martin.nv6@gmail.com

***Abstract*—El objetivo de este documento es desarrollar un modelo confiable de predicción de precios para propiedades de alquiler en Airbnb utilizando técnicas de aprendizaje automático. Se utilizarán características de alquileres y reseñas de clientes como predictores y se aplicarán una variedad de métodos, incluyendo regresión lineal, modelos basados en árboles, regresión Ridge, Lasso y Random Forest, para crear el modelo de predicción. El objetivo final es ayudar a propietarios y clientes a evaluar el precio óptimo de una propiedad con información limitada disponible.**

***Palabras clave*—** airbnb, predicción, precio, nueva york, regresión.

I. Introducción

La era de la pandemia de covid nos obligo a cambiar nuestra forma de vivir, con ello la manera en la que viajamos y nos hospedamos cambio, Airbnb es un servicio de interenet que provee la plataforma para conectar propietarios de casas o departamentos con clientes potenciales, durante la pandemia esta plataforma alcanzo niveles de uso muy altos, se realizo una comparación de exactitud de modelos de regresión para predicción de precios de hospedajes en Airbnb basado en una serie de parámetros como el tipo de habitación, la cantidad de opiniones en el sitio, cantidad de días de ocupación y el barrio.

A. Objetivo principal

Realizar una comparación de exactitud de modelos de regresión para predicción de precios de hospedajes en Airbnb.

B. Objetivo secundario

Con los resultados lograr encontrar un algoritmo de predicción de precios que pueda ser utilizado para calcular precios y encontrar las variables mas determinantes.

II. Antecedentes

Partes de la literatura existente sobre precios de propiedades se enfocan en la compra de propiedades no compartidas o predicciones de precios de alquiler. Previamente, Yu y Wu [1] intentaron implementar una predicción de precios de bienes raíces utilizando un análisis de importancia de características junto con regresión lineal, SVR y regresión Random Forest. También intentaron clasificar los precios en 7 clases utilizando Naive Bayes, Logistic Regression, SVC y Random Forest. Declararon un mejor RMSE de 0,53 para su modelo SVR y una precisión de clasificación del 69% para su modelo SVC con PCA. En otro artículo, Ma et al. [2] han aplicado Regresión Lineal, Árbol de Regresión, Regresión de Bosque Aleatorio y Árboles de Regresión de Aumento de Gradiente para analizar los precios de alquiler de almacenes en Beijing. Llegaron a la conclusión de que el modelo de regresión de árbol fue el modelo de mejor rendimiento con un RMSE de 1,05 CNY/m2-día. Otra clase de estudios, que son más pertinentes para este trabajo, inspeccionan los hoteles y los precios de alquiler de la economía compartida. En un trabajo reciente, Wang y Nicolau [3] estudiaron los determinantes de los precios de la economía colaborativa mediante el análisis de listados de Airbnb utilizando mínimos cuadrados ordinarios y análisis de regresión por cuantiles.

III. Descripción de datos

El conjunto de datos públicos de Airbnb para la ciudad de Nueva York se utilizó como fuente de datos principal para este estudio. El conjunto de datos incluía 50.221 entradas, cada una con 96 funciones. La figura 1 muestra la distribución geográfica de los precios de cotización en este conjunto de datos.

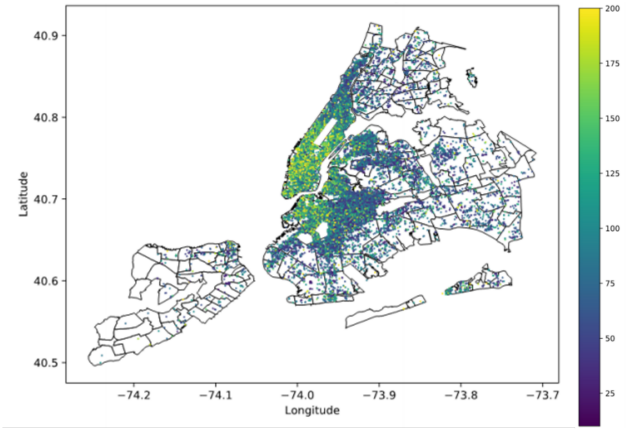
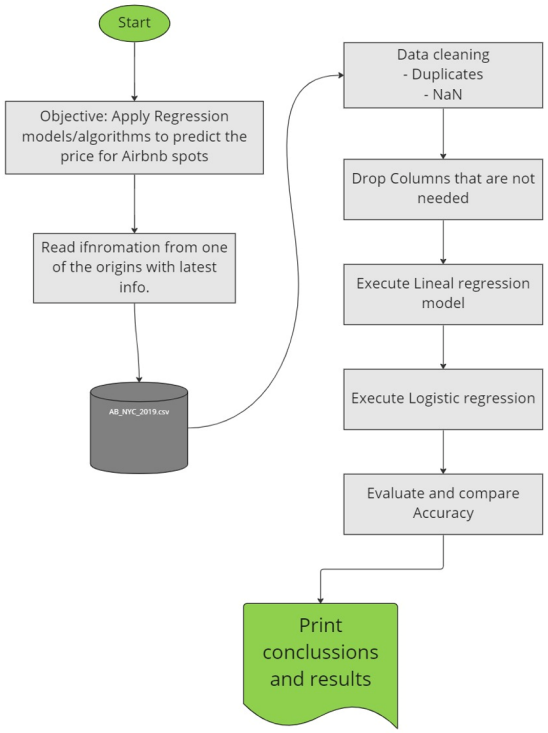


Figure 1: Geographic spread of price labels (with filtered outliers)

IV. Metodología

En el siguiente diagrama se muestra la metodología que se siguió para esta investigación



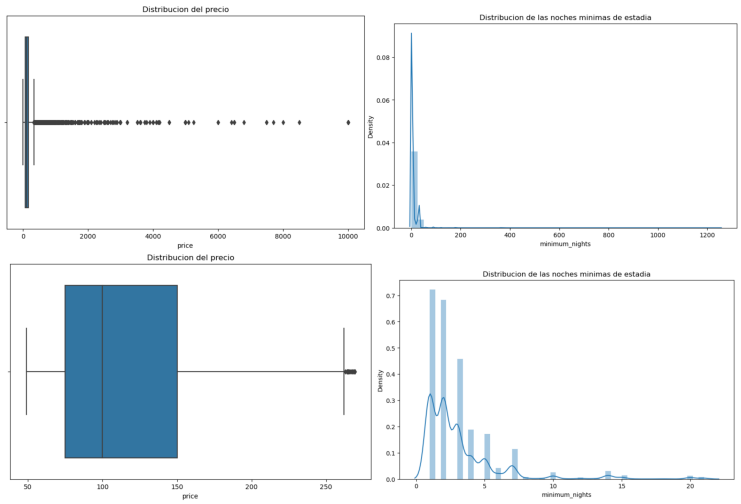
Se comienza con lectura de los datos y con limpieza de estos, para ello se hace lo siguiente

1) Remover duplicados
2) Encontrar los valores nulos
3) Reparar los valores nulos
4) En el caso de los valores nulos de 'reviewsPerMonth' se realiza un cambio de nulos por cero
5) Algunos campos son eliminados como el name, id, host name, last review, reviews per month, calculated host listings count
6) Se limpian los outliers [Figura2]
7) Se generan dummies para conversión de features categóricas a ficticias [Figura3]
8) Regresión lineal
   a) La regresión lineal que utiliza todo el conjunto de características como entradas del modelo se tomó como modelo de referencia para evaluando el desempeño de los otros métodos.
9) Regresión Ridge
   a) La regresión lineal con regularización L2 agrega un término de penalización a la función de costo del error cuadrático en para ayudar al algoritmo a converger para datos linealmente separables y reducir el sobre ajuste.
   b) Dado que se observó que los modelos de referencia tenían una varianza alta, Ridge Regression parecía ser una opción adecuada para resolver el problema.
10) Regresión Laso
   a) Tomando como consideración los resultados obtenidos en Regresion Ridge y lineal se decidió continuar con la ejecución de los siguientes modelos, con la intención de realizar suficientes comparaciones de resultados.
11) Regresión de Árbol de Decisión
   a) A pesar que se preveía que no seria un modelo útil se ejecuto par fines comparativos
12) Regresión Random Forest
   a) A pesar que se preveía que no seria un modelo útil se ejecuto par fines comparativos

V. Resultados

Como parte de los resultados de la eliminación de outlieres concluimos con lo que muestra la siguiente figura [figura2]

Figura 2: Eliminando outliers



Los dummies generados se muestran en la siguiente figura [figura3]

	Bronx	Brooklyn	Manhattan	Queens	Staten Island
0	0	1	0	0	0
1	0	0	1	0	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	1	0	0

En la [tabla 1] se muestran las métricas de rendimiento y desempeño de los modelos Desviación cuadrática media (RMSE) y R2 sirvieron para evaluar la modelos entrenados.

Tabla 1: Resultados de rendimiento

Modelo	R2 score (Entrenamiento)	R2 score (Prueba)	RMSE
Regresion Lineal	0.44	0.451	40.587
Regresion Ridge	0.44	0.451	40.578
Regresion Laso	0.214	0.219	40.578
Regresion de Arbol de Decission	0.813	0.289	40.578
Regresion Random Forest	1	0.101	40.578

VI. Conclusiones

Regresión Ridge y Regresión lineal tenían R2 similar y aunque el último modelo es más complejos que la regresión lineal. Se cree que la complejidad de la arquitectura de una red neuronal estaría limitada por la cantidad insuficiente de ejemplos de entrenamiento por tener demasiados pesos desconocidos.

Los datos indican que la Regresión de Laso no cuenta con puntuación suficiente como para ser considerado un modelo óptimo. Los resultados para Random forest y arbol de decisiones son muy ambiguos y no nos indican si el modelo es óptimo, además que el error de desviación cuadrática (RMSE) es alto llegando a 40.578% como se muestra en [tabla1]

VII. Trabajo a futuro

Debido a que los resultados no fueron del todo satisfactorios en términos de precisión por encima del 50%, como trabajo a futuro se realizara el procedimiento pero usando algoritmos de clasificación, de esta manera quizá se logre un mejor resultados buscando precios ”bajos”, ”medios” y ”altos”.

References

[1] H. Yu and J. Wu, “Real estate price prediction with regression and classification,” 2016. Accessed: 2023-3-18.

[2] D. Wang and J. L. Nicolau, “Price determinants of sharing economy based accommodationrental: A study of listings from 33 cities on airbnb. com,” 2018. Accessed: 2023-3-18.

[3] L. Schwarzová, Y. Satsangi, and E. Vanmassenhove, “Predicting airbnb prices with neighborhood characteristics: Machine learning approach.” Accessed: 2023-3-18.

[4] L. Schwarzová, “Predicting airbnb prices with neighborhood characteristics: Machine learning approach,” Dec. 2020. Accessed: 2023-3-18.

[5] P. R. Kalehbasti, L. Nikolenko, and H. Rezaei, “Airbnb price prediction using machine learning andsentiment analysis,” 2020. Accessed: 2023-3-18.

[1], [2], [3], [4], [5]