Proyecto Modelos y Simulación de Sistemas I

Entrega Final

Laura Victoria Ramos Agudelo

Profesor

Raúl Ramos Pollán, Professor of Computer Science

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería de sistemas

Medellín

Introducción

Planteamiento del Problema

La crisis de la vivienda holandesa es uno de los mayores problemas a los que se enfrentan los residentes. Debido a múltiples factores, como el crecimiento de la población y la escasez de trabajadores de la construcción, la disponibilidad de viviendas ha disminuido significativamente. Esta disminución ha llevado el alquiler a precios altísimos, lo que hace que muchos se pregunten si se están aprovechando de ellos.

Para responder a esta pregunta, el modelo debe predecir el alquiler de una casa a partir de sus características (es decir, ubicación, tamaño, instalaciones, etc.). El propósito de estos datos es poder investigar tendencias y patrones en el mercado de alquiler de bienes raíces en los Países Bajos. Con suerte, estos datos pueden explicar las situaciones actuales y ayudar a comprender lo que sucederá en este mercado.

Dataset o Base de Datos

El Dataset seleccionado es de una competición de Kaggle llamada Netherlands Accommodation Prices (FCG) y puede consultar en el siguiente enlace: https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/data. Esta base de datos contiene toda la información disponible en https://kamernet.nl/ para cada propiedad. El sitio web era rastreado diariamente y si aparecía una nueva propiedad, se añadía a la base de datos. Si se encontraba una propiedad que ya existía, se añadía a las fechas de publicación.

Métrica de Evaluación del Negocio

El modelo de predicción desarrollado ofrece la capacidad de abordar diversas preguntas. Puede identificar la zona más costosa para vivir, analizar qué tipo de propiedad proporciona la mejor relación calidad-precio y determinar cuál característica tiene el mayor impacto en el precio. Su utilidad se extiende a diferentes tipos de negocios. Las empresas dedicadas al alquiler de propiedades pueden utilizar este modelo para establecer precios óptimos. Del mismo modo, los inquilinos o empresas especializadas en encontrar viviendas para diversos públicos, como expatriados, pueden buscar propiedades rentables desviándose de este modelo.

En el contexto de nuestro proyecto, la métrica clave para evaluar el éxito del negocio se centra en el aumento de los ingresos operacionales de la empresa que nos ha contratado, la cual se dedica al alquiler de propiedades. El modelo proporciona la capacidad de determinar precios óptimos para las propiedades en alquiler.

Desempeño Deseable en Producción

Según el departamento de marketing de FCG, se espera que el modelo alcance un porcentaje de acierto de al menos el 70%. Este umbral es esencial, ya que el modelo se utilizará para determinar precios óptimos, y un porcentaje de acierto inferior podría resultar contraproducente al no compensar los costos asociados con mantener el modelo en producción.

Exploración Descriptiva del Dataset

Análisis Exploratorio y Preprocesado

El análisis exploratorio de los datos revela que el conjunto consta de un total de 27,915 filas y 34 columnas. Después de cambiar las columnas "firstSeenAt" y "lastSeenAt" a formato datetime tenemos que los tipos de nuestras columnas son: object, float64, int64 y datetime con 27, 3, 2 y 2 columnas respectivamente.

Debido a la limitada cantidad de información numérica disponible y la presencia de columnas con datos en formato de texto desafiante de procesar, como enlaces de imágenes o descripciones de publicaciones en línea, se ha optado por crear columnas adicionales que columnas adicionales para enriquecer los datos y posiblemente influir en la variable objetivo de la siguiente manera:

Precio de Alquiler por Metro Cuadrado (rent_per_areasqm): Se calculó el precio de alquiler por metro cuadrado dividiendo el precio de alquiler ("rent") por la superficie en metros cuadrados ("areaSqm"). Esta métrica ayuda a comprender la relación entre el precio de alquiler y el tamaño de la propiedad.

Distancia a las Ciudades Principales: Se calculó la distancia desde cada propiedad a las ciudades principales, lo que proporciona información valiosa sobre la ubicación de las propiedades en relación con los centros urbanos. Para este cálculo, se identificaron las 10 ciudades más frecuentemente listadas en

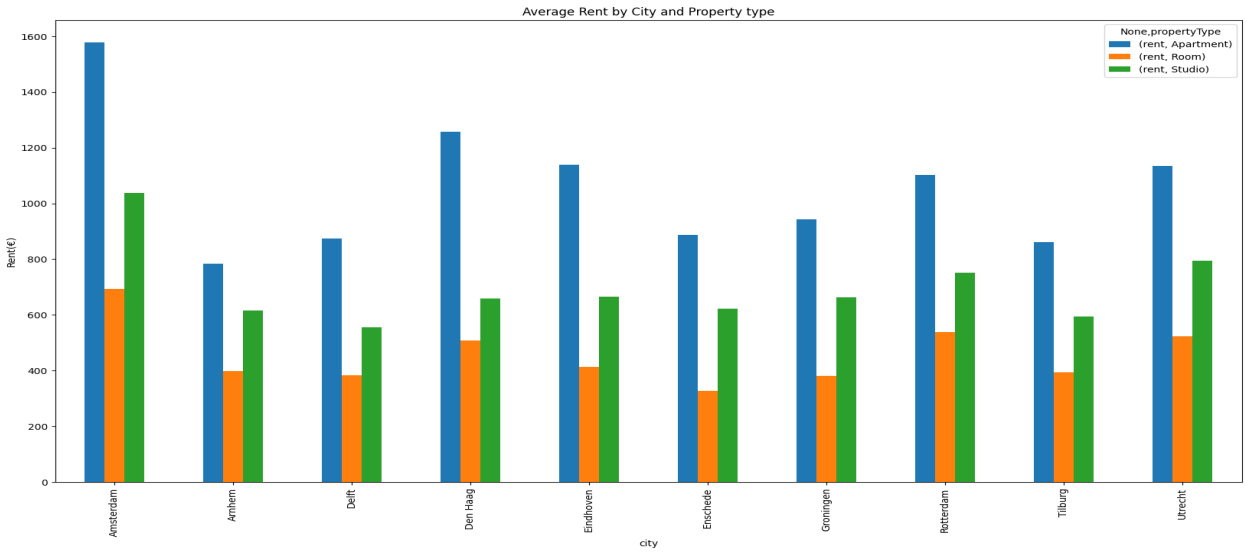
city	propertyType	count	rent	rent_per_areasqm
Amsterdam	Apartment	1358	1578.620029	25.315516
	Room	3124	693.408131	50.505142
	Studio	371	1037.247978	37.343747
Arnhem	Apartment	262	784.083969	15.730438
	Room	528	398.659091	25.361136
	Studio	54	615.833333	22.678839
Delft	Apartment	27	874.037037	17.083956
	Room	655	383.911450	26.322397
	Studio	25	556.040000	20.164271
Den Haag	Apartment	389	1257.442159	18.461863
	Room	841	509.122473	33.403859
	Studio	84	659.273810	23.405312

el conjunto de datos y se definió un diccionario llamado "city_centers" que contenía las coordenadas geográficas (latitud y longitud) de la estación central de cada una de estas ciudades. Luego, se aplicó la fórmula de haversine. para calcular la distancia en kilómetros desde cada propiedad, hasta el centro de la ciudad correspondiente.

Por otro lado, se creó un nuevo DataFrame llamado rent_city_property_type que contiene estadísticas resumidas sobre el alquiler en las ciudades grandes, desglosadas por el tipo de propiedad.

El DataFrame tiene índices multinivel con las categorías de "city" y "propertyType", y columnas que muestran el conteo, el promedio del alquiler y el promedio del alquiler por metro cuadrado para cada categoría. Esto permite un análisis detallado de cómo varía el alquiler en función de la ciudad y el tipo de propiedad

Con dicha información se creó un gráfico de barras que muestra el promedio del alquiler en diferentes ciudades y tipos de propiedad. Los valores de alquiler se muestran en el eje vertical (Y), y las barras representan el promedio del alquiler para cada combinación de ciudad y tipo de propiedad. Esto proporciona una representación visual de cómo varía el alquiler en función de la ciudad y el tipo de propiedad.

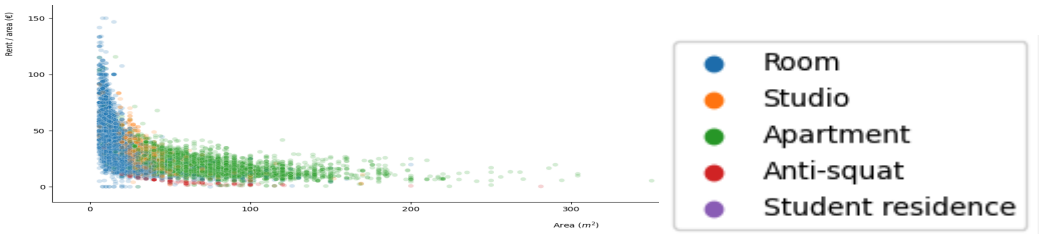


A partir de esta tabla y gráfico: si alguien desea vivir en un apartamento, Ámsterdam es la ciudad más cara con un promedio de 1600€ al mes, seguida de Den Haag con 1250€ al mes y Eindhoven, Rotterdam y Utrecht con alrededor de 1100€ al mes.

Para una habitación, el precio es de 700€ al mes en Ámsterdam, pero el precio por metro cuadrado será el más alto, a 50€ por metro cuadrado.

Para explorar la relación entre el tamaño, el precio y el tipo de propiedad, se utilizaron gráficos de dispersión que proporcionan una visualización efectiva de estos aspectos clave.

Relación entre Tamaño y Precio de Alquiler por Tipo de Propiedad



Las habitaciones, aunque pequeñas, son la opción más asequible. En contraste, los apartamentos varían en tamaño y precio, lo que sugiere una mayor diversidad de opciones. Es evidente que los apartamentos pueden presentar una mayor variabilidad en términos de precio y tamaño en comparación con las habitaciones.

Al observar los gráficos, es claro que a medida que el tamaño de la propiedad aumenta, también lo hace el precio. La relación entre el tamaño y el precio es directa. Sin embargo, existe una excepción en el caso de las propiedades "Anti-squat", que mantienen un precio constante independientemente del tamaño.

Variables Categóricas

Al inspeccionar las variables categóricas, surgen varias preocupaciones sobre cómo procesar algunas de las columnas, como "matchAge," "matchLanguages," "coverImageUrl," y "descriptionNonTranslated." Estas columnas presentan una amplia variedad de categorías y valores únicos, lo que podría complicar su inclusión en un modelo analítico. En el caso de "matchAge," encontramos numerosas categorías que representan rangos de edades, pero también algunas categorías como "Not important - Not important." La diversidad de categorías podría dificultar la interpretación y el análisis. Con respecto a "matchLanguages," hay una gran cantidad de combinaciones de idiomas, lo que hace que esta columna sea compleja de manejar y puede requerir una codificación especial. Y lo mismo sucede con otras columnas, lo que plantea desafíos adicionales para su procesamiento en un modelo analítico.

Limpieza de Datos

Dada la complejidad previamente mencionada con algunas variables categóricas, se llevó a cabo una preselección de columnas, optando por aquellas que fueran

numéricas o categóricas con menos de 7 valores únicos. De esta preselección resultaron 33 columnas. Posteriormente, se procedió a eliminar la columna 'id', que no aporta información relevante, así como las columnas 'latitude' y 'longitude', que se utilizaron para el cálculo de las distancias a las principales ciudades. Además, se excluyó 'rentDetail' por contener solo un valor único: 'Utilities incl.'.

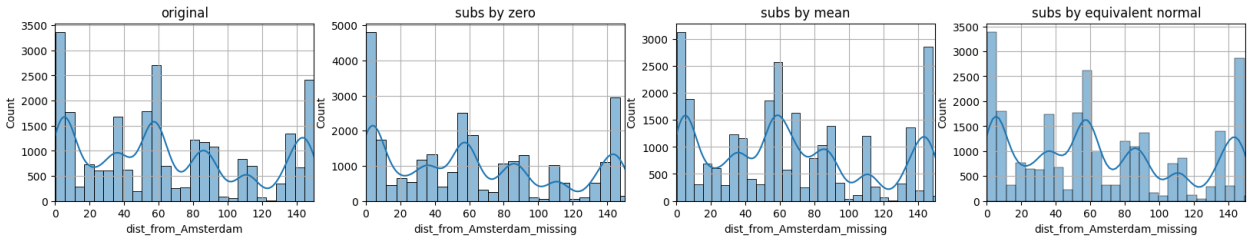
Se identificó que alquilar un 'Anti-squat' tiene un precio constante, independientemente del tamaño de la propiedad. Sin embargo, dado que "Anti-squat" es una residencia temporal y no está vinculada a ninguna otra propiedad, se ha decidido eliminar estos datos para mejorar la integridad del modelo.

Reparar Datos Faltantes

La base de datos es muy completa, el mayor porcentaje de datos faltantes es de 2% en las columnas ‘gender’ y ‘roommates’. Teniendo en cuenta que entre las condiciones del presente proyecto se encuentra que el conjunto de datos “ha de tener un 5% de datos faltantes en al menos 3 columnas.” , por lo cuál se hizo una selección aleatoria de 3 columnas una única vez y a estas se le eliminaron aleatoriamente el 5% de los datos.

Para completar los datos de las variables categóricas se usó la estrategia de reemplazarlas con el dato más frecuente. Para reparar los datos de la única variable numérica con faltantes se realizó una comparación visual de diferentes estrategias de imputación de valores faltantes generando tres versiones adicionales del DataFrame original, cada una utilizando una estrategia diferente para imputar los valores faltantes en ‘dist_from_Amsterdam’. Estas estrategias incluyen sustituir los

valores faltantes por cero, por la media de la columna, y por valores generados aleatoriamente siguiendo una distribución normal equivalente a la columna original.



La técnica de sustitución más adecuada en este caso parece ser la generación de valores aleatorios siguiendo una distribución normal equivalente a la columna original. Es importante señalar que esta conclusión puede no ser definitiva y requiere una validación adicional mediante el modelado, considerando las tres técnicas de imputación. No obstante, para los propósitos de este ejercicio académico, se optó por esta decisión, siendo consciente de que podría impactar los resultados del modelo seleccionado.

Iteraciones de desarrollo

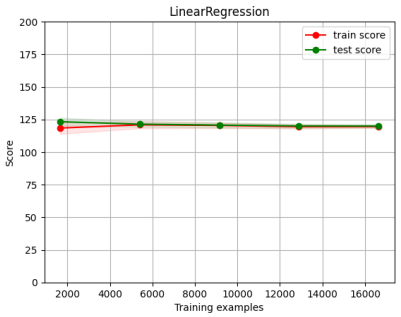
Es fundamental destacar que la métrica de evaluación de Machine Learning para el modelo será el Error Absoluto Promedio (MAE), el cual proporcionará el promedio de la diferencia absoluta entre las predicciones del modelo y los valores objetivos. Un alto Mean Absolute Error (MAE) en el contexto de un modelo de regresión o predicción indica que las predicciones generadas por el modelo tienden a estar significativamente alejadas de los valores reales o de referencia.

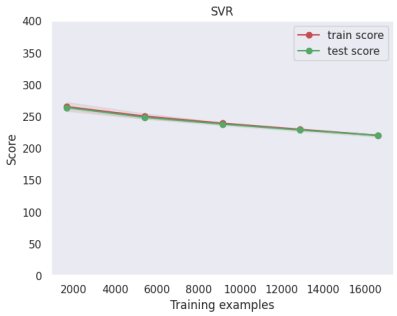
Con esto en mente, tras realizar One-Hot Encoding, se dividió el conjunto de datos en un 20% para evaluar la eficiencia del modelo y un 80% para entrenarlo. En la primera iteración, se emplearon cuatro modelos supervisados: Regresión Lineal, Support Vector Machine (SVR) con gamma=1, Modelo de Árbol de Decisión y Modelo de Bosque Aleatorio. Estos dos últimos se configuraron con una profundidad máxima de 5. Los resultados de esta primera iteración fueron los siguientes:

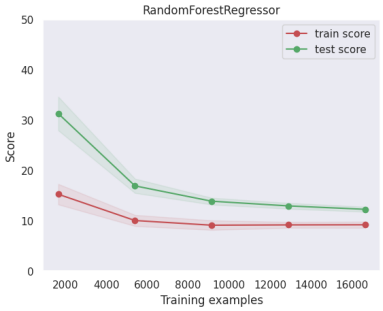
- Regresión Lineal MAE: 117.4753. test_score: 0.783726 train_score: 0.76820.
- SVR MAE: 282.5392. test_score: -0.08515 train_score: -0.08048.
- Árbol de Decisión MAE: 110.7174. test_score: 0.85422 train_score: 0.8734951.
- Random Forest MAE: 89.5463. test_score: 0.89812 train_score: 0.911083.

Siendo el modelo Random Forest el que arroja mejores resultados en esta iteración.

En la segunda iteración se realizó Bootstrapping con 10 divisiones diferentes de la data probando los mismos 4 modelos, para el SVR se usó un gamma igual a ‘scale’ y para Decision Tree Model y Random Forest Model se usó una profundidad máxima de 10. En esta iteración se descarta el uso de SVR debido a su ejecución extremadamente demorada y a sus resultados poco satisfactorios. El mejor modelo para esta iteración nuevamente es Random Forest que arrojó un test score de 15.702 (±1.5151), siendo el menor de todos. También se hicieron gráficas de las curvas de aprendizaje de cada modelo:







Con cierta incertidumbre o limitado conocimiento sobre la interpretación de comportamientos de modelos a través de curvas de aprendizaje, se puede sugerir que el desempeño del modelo RandomForest es el esperado. La representación visual del modelo con árboles de decisión parece indicar un posible problema de sesgo (bias), aunque es importante señalar que esta gráfica no se considera concluyente para efectos de este proyecto.

Feature	Importance
areaSqm	0.699621
rent_per_areasqm	0.221119
dist_from_Amsterdam	0.041982
propertyType_Room	0.024444
dist_from_DenHaag	0.009458
dist_from_Eindhoven	0.000392
dist_from_Groningen	0.000379
dist_from_Enschede	0.000238
dist_from_Delft	0.000234
dist_from_Arnhem	0.000225

En la tercera iteración, se opta por identificar las características más relevantes para predecir el arrendamiento de propiedades en Países Bajos utilizando el estimador RandomForest. Los resultados respaldan las observaciones de la exploración descriptiva realizada inicialmente, destacando que el área de la propiedad y la distancia a las ciudades principales son los factores que mejor explican el valor de la variable objetivo. Con este conocimiento, se decide modelar utilizando sólo estas variables, empleando tres estimadores (regresión lineal, árboles de decisión y Random Forest). Se modifica la profundidad máxima de los dos últimos a 20. Una vez más, se concluye que el mejor modelo es el Random Forest, mejorando sus predicciones con un test

score de 7.518 (±0.4292). Se plantea realizar una última iteración utilizando todas las columnas seleccionadas inicialmente y la misma profundidad de 20 para determinar si la mejora del modelo se debe a la nueva selección de columnas o al aumento de la profundidad.

La última iteración, la cuarta, revela que el mejor modelo sigue siendo el Random Forest, con un test score de 7.636 (±0.8116), ligeramente superior al modelo anterior. La elección del modelo se torna más compleja debido al rendimiento similar y destacado de ambos modelos. En el contexto de este proyecto, se opta por el RandomForestRegressor con parámetro max_depth=20 y utilizando la data depurada de la iteración tres. Esta elección se fundamenta en su mayor confiabilidad según los resultados de MAE y en la eficiencia. A continuación, se listan las ventajas que se tienen en cuenta para elegir este modelo sobre el mismo realizado con todas las variables elegidas inicialmente:

Eficiencia Computacional: Al reducir el número de columnas, el modelo requiere menos recursos computacionales para entrenar y predecir, lo que se traduce en un proceso más eficiente.

Interpretación y Mantenimiento: Un modelo con menos columnas es más fácil de interpretar y mantener. Se simplifica la comprensión de las características más relevantes, facilitando la identificación y corrección de posibles problemas o mejoras.

Ahorro de Recursos: Al requerir menos datos, se ahorra tiempo y esfuerzo en la recopilación, limpieza y procesamiento de información, lo cual es especialmente valioso en proyectos donde la eficiencia es clave.

Enfocar en Características Relevantes: La selección cuidadosa de las características más relevantes, como el área de la propiedad y la distancia a las ciudades principales, permite concentrarse en factores que tienen un impacto significativo en la variable objetivo, mejorando así la capacidad explicativa del modelo.

Eficiencia del Modelo

El 20% de los datos iniciales, reservados exclusivamente para este paso del proceso, fueron transformados de manera que sólo incluyeran las columnas relevantes, siguiendo el mismo enfoque utilizado en la iteración número 3, de la cual se derivó el modelo seleccionado. Posteriormente, el modelo se entrenó con el restante 80% de los datos, y se evaluó su rendimiento utilizando los nuevos datos, lo que resultó en un resultado muy positivo, ya que el Error Absoluto Medio (MAE) disminuyó a 6.321.

Según el departamento de marketing de FCG, un modelo de predicción de precios de alquiler de propiedades en los Países Bajos debería tener un porcentaje de acierto de al menos el 70%. Si el porcentaje de acierto es inferior, sería contraproducente, ya que los ingresos no compensarán los costos de mantener el modelo en producción. Al calcular el Error Absoluto Medio Relativo (EAMR) en promedio, se encontró que las predicciones del modelo tienen un error relativo del 4.29%. En otras palabras, la magnitud promedio de la diferencia entre las predicciones del modelo y los valores reales es aproximadamente el 4.29%, lo que indica que el modelo está cumpliendo con el rendimiento esperado y puede mantenerse en producción.

Retos y Consideraciones de Despliegue

Complejidad en Variables Categóricas

El manejo de variables categóricas, como "matchAge," "matchLanguages," "coverImageUrl," y "descriptionNonTranslated," presenta desafíos debido a la diversidad de categorías y valores únicos. La codificación y procesamiento de estas columnas para su inclusión en un modelo analítico son tareas que requieren atención especial.

Lidiar con Datos Faltantes en Variables Numéricas

La identificación y tratamiento de datos faltantes en variables numéricas, especialmente la columna 'dist_from_Amsterdam', implica la selección de estrategias adecuadas para imputar estos valores y garantizar la integridad del modelo.

Selección de Variables Relevantes

La selección cuidadosa de variables relevantes para el modelo representa un reto, ya que la elección incorrecta de características puede afectar la capacidad explicativa del modelo y, por ende, su rendimiento predictivo.

Sesgo en Modelos de Regresión

La interpretación de las curvas de aprendizaje y la identificación de posibles problemas de sesgo (bias) en modelos de regresión, como el RandomForest, requieren comprensión y validación adicionales para tomar decisiones informadas.

Tiempo de Ejecución en Support Vector Machine (SVM)

La ejecución del modelo Support Vector Machine (SVM) lleva aproximadamente 12 minutos por cada iteración, lo que representa un desafío en términos de eficiencia computacional. La duración prolongada puede afectar la agilidad del proceso de modelado y exploración de hiperparámetros.

Conclusiones

Eficiencia del Modelo Seleccionado

La elección del modelo RandomForestRegressor con parámetro max_depth=20 y utilizando datos depurados en la tercera iteración se basa en su mayor confiabilidad según los resultados de MAE y en la eficiencia computacional. Este modelo destaca por su capacidad para predecir el precio de alquiler con precisión y requerir menos recursos computacionales.

Validación con Métricas de Evaluación

El modelo seleccionado ha superado las expectativas del departamento de marketing de FCG, con un Error Absoluto Medio (MAE) reducido a 6.321 y un Error Absoluto Medio Relativo (EAMR) promedio del 4.29%. Estos resultados cumplen con los criterios establecidos para el porcentaje de acierto y respaldan la decisión de mantener el modelo en producción.

Optimización de Recursos

La reducción del conjunto de características a las más relevantes ha demostrado ser ventajosa en términos de eficiencia computacional. La elección de características específicas, como el área de la propiedad y la distancia a las ciudades principales, no solo mejora la capacidad explicativa del modelo sino que también simplifica su interpretación y mantenimiento.

Desafíos Futuros

Aunque el modelo ha demostrado un rendimiento sólido, la gestión de variables categóricas y la selección de estrategias para manejar datos faltantes siguen siendo áreas que podrían beneficiarse de un enfoque más detallado en futuros proyectos de análisis y modelado.

Cita	Netherlands Accommodation Prices (FCG)
Referencia	[1] Kaggle. (s. f.). https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview
Estilo (2020)	IEEE