|  |  |
| --- | --- |
|  |  |
| Proyecto Supervisado. Técnicas de aprendizaje de máquina. |  |
| Pontificia Universidad Javeriana - Wikipedia |  |
|  | 22/04/2024Predicción precios viviendas. Entregables. |
|  | Juan Esteban Rivera y Victoria Chavarro |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | *Limpieza y manipulación de los datos* Durante la fase de limpieza y manipulación de datos, se llevaron a cabo una serie de etapas para preparar los datos para su análisis y modelado. A continuación, se describe detalladamente cada una de estas etapas:1. Carga de Datos:En esta etapa, se procedió a cargar los conjuntos de datos proporcionados, que incluyen información sobre características de los inmuebles y sus precios totales, así como datos complementarios como puntos de interés geográfico. Se utilizó el formato adecuado para leer los datos y se verificó la integridad de la carga para asegurar que todos los registros se importaron correctamente.2. Eliminación Inicial de Variables Innecesarias:Se realizó una revisión inicial de las variables presentes en los conjuntos de datos para identificar aquellas que no aportan valor al análisis o modelado. Se eliminaron las variables que no estaban relacionadas con el objetivo del proyecto o que tenían una baja correlación con la variable objetivo (valor\_total\_avaluo). Esto permitió simplificar el conjunto de datos y reducir la complejidad del análisis.3. Manejo de Nulos Inicial:Se llevó a cabo una evaluación preliminar de los valores nulos en los datos y se implementaron estrategias iniciales para manejar estos valores. Esto incluyó técnicas como la imputación de valores utilizando la media o la mediana de la columna correspondiente, o la eliminación de registros con valores nulos, dependiendo del impacto en la calidad de los datos y la cantidad de registros afectados.4. Transformación de Columnas/Variables:Se realizaron transformaciones en las columnas o variables pertinentes para adaptarlas al formato adecuado requerido por los algoritmos de modelado. Esto incluyó la conversión de variables categóricas a numéricas utilizando técnicas como la codificación one-hot o la asignación de valores numéricos a las categorías. También se llevaron a cabo transformaciones para normalizar los datos numéricos y reducir la escala de las variables.5. Corrección de Caracteres Especiales:Se identificaron y corrigieron caracteres especiales o inconsistencias en los datos que podrían afectar su procesamiento o interpretación. Se utilizaron técnicas de limpieza de texto para eliminar caracteres no deseados o convertirlos en un formato legible y coherente. *6. Transformación de Variables Descripción/Observación:*  *Durante esta etapa, se exploraron variables que contenían información descriptiva u observacional, como descripciones de propiedades o notas adicionales. Se realizaron transformaciones para extraer características relevantes o convertirlas en variables categóricas utilizables en el modelado.*  *7. Preprocesamiento de Variables Categóricas:*  *Las variables categóricas fueron preprocesadas para convertirlas en un formato adecuado para su inclusión en los modelos de aprendizaje automático. Esto involucró la selección de variables que queríamos convertir en categóricas, sus valores únicos, etc. Este paso fue crucial para garantizar que las variables categóricas pudieran ser interpretadas correctamente por los algoritmos de modelado.*  *8. Manejo de Valores Corridos/Incoherentes:*  *Se identificaron y corrigieron valores corridos o incoherentes en los datos que podrían ser el resultado de errores de entrada o medidas extremas atípicas. Se aplicaron técnicas como la detección de valores atípicos utilizando métodos estadísticos o visuales, y se tomaron medidas adecuadas según el contexto de los datos y los requisitos del proyecto, como la eliminación de valores atípicos o la imputación de valores basada en criterios específicos.*  *9. Conversión de Variables a Categóricas:*  *En esta etapa, se identificaron variables que originalmente estaban en formato numérico pero que representaban categorías discretas. Estas variables fueron convertidas al formato categórico correspondiente, ya que su naturaleza no era puramente numérica. Por ejemplo, variables como "tipo de propiedad" o "tipo estructura" podrían ser convertidas a variables categóricas para un mejor manejo y comprensión en el análisis.*  *10. Manejo de Variables Binarias:*  *Se realizaron ajustes en las variables para asegurar su coherencia y relevancia en el análisis. Esto podría incluir la verificación de la consistencia en la codificación de variables binarias, asegurando que reflejen adecuadamente la presencia o ausencia de ciertas características, y la eliminación de variables binarias redundantes o irrelevantes para el análisis.*  *11. Manejo de Variables Enteras:*  *Las variables de tipo entero fueron examinadas y ajustadas según fuera necesario para garantizar su adecuación al análisis. Esto podría incluir la detección y corrección de valores negativos no válidos, la normalización de rangos de valores para mejorar la comparabilidad entre variables, y la conversión de variables enteras a otros formatos según las necesidades del modelado, como variables binarias o categóricas.*  *12. Eliminación de Índices:*  *Se eliminaron los índices o identificadores únicos que no aportaban información relevante al análisis o modelado, ya que podrían introducir ruido o sesgo en los resultados. Esto se hizo para simplificar la estructura de los datos y reducir la complejidad en los modelos posteriores.*  *13. Otras Transformaciones:*  *Se llevaron a cabo otras transformaciones de datos según las necesidades específicas del proyecto, como la creación de nuevas variables derivadas a partir de combinaciones de variables existentes, la normalización de datos para mejorar la precisión de los modelos, o la conversión de formatos de datos para facilitar su manipulación y análisis.*  *14. Manejo de Variables Object y Float:*  *Las variables de tipo object y float fueron examinadas y ajustadas según fuera necesario para asegurar su adecuación al análisis. Esto podría incluir la conversión de variables object a formatos numéricos o categóricos o incluso su eliminación, y la estandarización de variables float para mejorar la comparabilidad entre diferentes características.*  *15. Análisis de Relación entre Variables Numéricas y Categóricas:*  *Se exploraron las relaciones entre variables numéricas y categóricas para identificar posibles patrones o correlaciones significativas que pudieran influir en el modelado. Se utilizaron técnicas de análisis estadístico y visualización de datos para examinar estas relaciones y entender mejor la estructura de los datos.*  *16. Estandarización de Variables Float:*  *Las variables float fueron estandarizadas para asegurar que todas las características tuvieran la misma escala y contribuyeran equitativamente al análisis. Esto se logró mediante técnicas como la normalización o la estandarización de datos, que ajustaron los valores de las variables para que estuvieran centrados alrededor de cero y tuvieran una desviación estándar de uno.*  *Estas etapas de limpieza y manipulación de datos fueron fundamentales para garantizar la calidad y la idoneidad de los datos para su posterior análisis y modelado. Cada paso contribuyó a la preparación efectiva de los datos y al desarrollo de modelos de alta calidad que pudieran generar resultados precisos y significativos para el proyecto.* *Exploración de Datos* *La exploración de datos es una fase crucial en cualquier proyecto de análisis y modelado, ya que proporciona una comprensión profunda de la estructura, la calidad y las relaciones entre las variables presentes en el conjunto de datos. Durante esta etapa, se llevaron a cabo diversas actividades para examinar en detalle la información disponible y establecer las bases para el análisis posterior. A continuación, se detallan las acciones realizadas durante la exploración de datos:*  *1. Análisis de Variables:*  *Se realizó un análisis exhaustivo de las variables presentes en el conjunto de datos, incluyendo tanto las variables predictoras como la variable objetivo (valor\_total\_avaluo). Se examinaron los tipos de variables (numéricas, categóricas, etc.), su distribución, rangos de valores y posibles valores atípicos o nulos.*  *2. Visualización de Datos:*  *Se utilizaron técnicas de visualización de datos para representar gráficamente la distribución de las variables y explorar posibles patrones o relaciones entre ellas. Se generaron histogramas, diagramas de dispersión, matrices de correlación y otros tipos de gráficos según la naturaleza de las variables y los objetivos del análisis.*  *3. Análisis de Correlación:*  *Se llevó a cabo un análisis de correlación entre las variables predictoras y la variable objetivo para identificar posibles relaciones significativas que pudieran influir en el modelado. Se utilizaron coeficientes de correlación y gráficos de dispersión para evaluar la fuerza y la dirección de estas relaciones.*  *4. Identificación de Valores Atípicos:*  *Se examinaron los valores extremos o atípicos en las variables para determinar su impacto en el análisis y el modelado. Se utilizaron técnicas estadísticas y visuales para detectar valores atípicos y se evaluó su influencia en los resultados obtenidos.*  *5. Manejo de Datos Faltantes:*  *Se exploraron los valores nulos o faltantes en el conjunto de datos y se evaluaron las estrategias para manejar estos valores de manera adecuada. Se examinaron las posibles causas de los valores faltantes y se implementaron técnicas de imputación o eliminación de registros según el contexto de los datos y los requisitos del análisis.*  *6. Exploración de Variables Descriptivas:*  *Se analizaron variables descriptivas u observacionales, como descripciones de propiedades o notas adicionales, para identificar posibles características relevantes o patrones ocultos. Se utilizaron técnicas de procesamiento de lenguaje natural (NLP) o análisis de texto para extraer información útil de estas variables y convertirlas en características numéricas para su inclusión en el modelado.*  *7. Evaluación de la Calidad de los Datos:*  *Se llevó a cabo una evaluación de la calidad de los datos para identificar posibles errores, inconsistencias o sesgos que pudieran afectar el análisis. Se examinaron la integridad de los registros, la consistencia entre variables y la presencia de datos duplicados o redundantes.*  *La exploración de datos proporcionó una visión completa de la estructura y la calidad de los datos, así como de las posibles relaciones entre variables. Esta etapa fue fundamental para guiar las decisiones posteriores de limpieza, manipulación y modelado de datos, y para garantizar la fiabilidad y la relevancia de los resultados obtenidos en el proyecto.* *Técnicas Usadas o Exploradas* *Durante el desarrollo del proyecto, se emplearon diversas técnicas de modelado y aprendizaje automático para abordar la estimación de precios de vivienda. A continuación, se explican detalladamente las técnicas utilizadas y su aplicación en el contexto del proyecto:*  *1. Máquinas de Vectores de Soporte (SVM):*  *Las Máquinas de Vectores de Soporte (SVM) son un algoritmo de aprendizaje supervisado que se utiliza principalmente para problemas de clasificación y regresión. En el contexto de este proyecto, se exploró la aplicación de SVM para la estimación de precios de vivienda como un modelo de regresión. SVM busca encontrar el hiperplano que mejor separa los datos de entrenamiento en el espacio de características, con el objetivo de minimizar el error de predicción en el conjunto de datos.*  *2. Red Neuronal:*  *Las redes neuronales son modelos computacionales inspirados en el funcionamiento del cerebro humano, que se utilizan en una amplia variedad de aplicaciones de aprendizaje automático. En este proyecto, se implementó una red neuronal para la estimación de precios de vivienda. La red neuronal consistió en múltiples capas de neuronas interconectadas, con una capa de entrada, una o más capas ocultas y una capa de salida. Se utilizó la retropropagación para ajustar los pesos de las conexiones y minimizar el error de predicción en el conjunto de entrenamiento.*  *3. Random Forest:*  *Random Forest es un algoritmo de aprendizaje supervisado que se utiliza principalmente para problemas de clasificación y regresión. En el contexto de este proyecto, se exploró la aplicación de Random Forest para la estimación de precios de vivienda como un modelo de regresión. Random Forest se basa en la construcción de múltiples árboles de decisión durante el entrenamiento, donde cada árbol se entrena con una muestra aleatoria del conjunto de datos y realiza predicciones individuales. Luego, se promedian las predicciones de todos los árboles para obtener una predicción final.*  *4. Regresión Lineal Predictora:*  *La regresión lineal es un modelo estadístico que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. En este proyecto, se implementó la regresión lineal predictora para estimar los precios de vivienda utilizando un enfoque lineal. La regresión lineal busca encontrar la mejor línea de ajuste que minimice la suma de los errores cuadráticos entre las predicciones del modelo y los valores reales en el conjunto de datos.*  *5. Árbol de Regresión:*  *Los árboles de regresión son modelos de aprendizaje supervisado que se utilizan para problemas de regresión, donde la variable objetivo es continua. En el contexto de este proyecto, se emplearon árboles de regresión para predecir los precios de vivienda. Los árboles de regresión dividen el espacio de características en regiones más pequeñas y realizan predicciones basadas en la media de los valores de las muestras de entrenamiento en cada región.*  *Cada una de estas técnicas fue explorada y evaluada en función de su rendimiento y capacidad para modelar de manera efectiva la relación entre las características de las viviendas y sus precios. La selección de la técnica final se basó en su capacidad para generar predicciones precisas y consistentes en el conjunto de datos de prueba.* *Técnica elegida* *La elección de la técnica adecuada para abordar el problema de estimación de precios de vivienda es un paso crítico en el proceso de modelado. Considerando las opciones disponibles, tanto la Regresión Lineal Predictora como el Árbol de Regresión presentan ventajas y consideraciones importantes que deben tenerse en cuenta. A continuación, se analizan ambas opciones:*  *Regresión Lineal Predictora:*  *La Regresión Lineal Predictora ofrece una aproximación inicialmente atractiva debido a su simplicidad y facilidad de interpretación. Este modelo asume una relación lineal entre las características de las viviendas y sus precios, lo cual puede ser razonable en ciertos contextos. Además, la Regresión Lineal Predictora es fácilmente interpretable y puede proporcionar una comprensión intuitiva de cómo cada característica contribuye al precio de la vivienda.*  *Sin embargo, al considerar la complejidad del problema de estimación de precios de vivienda, es importante reconocer que la relación entre las características de las viviendas y sus precios puede ser más compleja y no lineal. La Regresión Lineal Predictora puede no capturar eficazmente esta complejidad, lo que podría limitar su capacidad para realizar predicciones precisas en datos donde las relaciones son más intrincadas.*  *Al evaluar el rendimiento del modelo, se observó un coeficiente de determinación (R2) alto, lo que indica que una gran proporción de la variabilidad en los precios de las viviendas puede ser explicada por las características consideradas en el modelo. Aunque el valor de R2 para la Regresión Lineal Predictora fue ligeramente más alto que el del Árbol de Regresión, se debe tener en cuenta que este resultado puede deberse a la naturaleza simplista del modelo lineal.*  *Árbol de Regresión:*  *Por otro lado, el Árbol de Regresión presenta una alternativa más flexible y adaptable. Este modelo es capaz de capturar relaciones tanto lineales como no lineales entre las características de las viviendas y sus precios, gracias a su capacidad para dividir el espacio de características en segmentos más pequeños de manera recursiva.*  *El Árbol de Regresión es altamente interpretable, ya que divide el espacio de características en regiones más pequeñas y realiza predicciones basadas en reglas simples que pueden ser fácilmente visualizadas y entendidas. Esto proporciona una ventaja significativa en términos de comprensión del modelo y comunicación de los resultados a partes interesadas.*  *Al evaluar el rendimiento del Árbol de Regresión, se observó un coeficiente de determinación (R2) también alto, aunque ligeramente menor que el de la Regresión Lineal Predictora. Esto sugiere que el modelo es capaz de explicar una gran parte de la variabilidad en los precios de las viviendas, pero puede haber margen para mejorar su capacidad predictiva.*  *Elección Final:*  *Considerando las características del problema de estimación de precios de vivienda y la capacidad de los modelos para capturar relaciones complejas en los datos, el Árbol de Regresión emerge como una opción prometedora. Su capacidad para capturar tanto relaciones lineales como no lineales, junto con su interpretabilidad, hacen que sea una opción atractiva para este tipo de problema.*  *Sin embargo, es importante tener en cuenta que la Regresión Lineal Predictora también podría ofrecer resultados aceptables en ciertos casos, especialmente si la relación entre las características de las viviendas y sus precios es predominantemente lineal. Por lo tanto, aunque el Árbol de Regresión parece ser la mejor elección dadas las características del problema, se debe realizar una evaluación detallada de ambos modelos para determinar la opción más adecuada en función de la precisión de las predicciones y la interpretabilidad del modelo.*  *Además, se pueden explorar estrategias para mejorar el rendimiento del Árbol de Regresión, como considerar técnicas de ensemble para combinar múltiples árboles y mejorar la generalización del modelo.* *Dificultades Encontradas* *Durante el desarrollo del proyecto de estimación de precios de vivienda, se enfrentaron diversas dificultades que requirieron atención y resolución. Estas dificultades abarcaron desde desafíos técnicos hasta limitaciones en los datos disponibles, y se abordaron con estrategias específicas para garantizar el éxito del proyecto. A continuación, se detallan las principales dificultades encontradas:*  *1. Calidad de los Datos:*  *Uno de los principales desafíos encontrados fue la calidad de los datos disponibles. Se identificaron problemas como valores faltantes, valores atípicos o errores de entrada que podrían afectar la precisión de los modelos de estimación de precios de vivienda. Para abordar este desafío, se implementaron técnicas de limpieza de datos, imputación de valores faltantes y detección y corrección de valores atípicos para garantizar la integridad y la consistencia de los datos utilizados en el análisis.*  *2. Complejidad del Problema:*  *El problema de estimación de precios de vivienda es inherentemente complejo debido a la amplia variedad de factores que pueden influir en el precio de una propiedad. La identificación y la captura precisa de todas estas influencias pueden ser desafiantes y requieren un análisis detallado de las características de las viviendas y su entorno. Para abordar esta complejidad, se llevaron a cabo exploraciones exhaustivas de datos y se consideraron múltiples enfoques de modelado para capturar la mayor cantidad posible de información relevante.*  *3. Interpretación de Modelos:*  *Otra dificultad encontrada fue la interpretación de los modelos de estimación de precios de vivienda, especialmente en el caso de modelos más complejos como los árboles de regresión. La comprensión de cómo las características de las viviendas influyen en el precio puede ser crucial para tomar decisiones informadas, pero la interpretabilidad de algunos modelos puede ser limitada. Se exploraron técnicas de visualización y análisis de modelos para abordar este desafío y proporcionar una comprensión más clara de los factores que influyen en las predicciones del modelo.*  *4. Optimización del Rendimiento del Modelo:*  *La optimización del rendimiento de los modelos de estimación de precios de vivienda fue otro desafío importante. Aunque se obtuvieron resultados prometedores con los modelos implementados, existía la necesidad de mejorar aún más la precisión de las predicciones y la capacidad de generalización de los modelos. Se llevaron a cabo experimentos con ajustes de hiperparámetros, técnicas de selección de características y enfoques de ensemble para mejorar el rendimiento de los modelos y garantizar que fueran lo más precisos y robustos posible.* *Aspectos Importantes* *1. Validación Cruzada: Uno de los aspectos importantes en el desarrollo del proyecto fue la implementación de técnicas de validación cruzada para evaluar el rendimiento de los modelos de estimación de precios de vivienda. La validación cruzada permitió una evaluación más robusta de los modelos al dividir el conjunto de datos en múltiples particiones y realizar múltiples iteraciones de entrenamiento y evaluación. Esto ayudó a mitigar el riesgo de sobreajuste y proporcionó una estimación más precisa del rendimiento esperado del modelo en datos no vistos.*  *2. Selección de Características: Otro aspecto crucial fue la selección de características o variables más relevantes para la estimación de precios de vivienda. Se realizaron análisis detallados de la importancia de las características y se exploraron técnicas de selección de características para identificar y utilizar únicamente aquellas que contribuyeran significativamente a la precisión del modelo. Esto permitió simplificar los modelos y mejorar su capacidad de generalización al eliminar características redundantes o poco informativas.*  *3. Interpretación de Resultados: La interpretación de los resultados obtenidos de los modelos de estimación de precios de vivienda fue un aspecto fundamental para obtener insights significativos y tomar decisiones informadas. Se prestaron especial atención a las características que más influyeron en las predicciones del modelo, así como a posibles patrones o tendencias identificados durante el análisis. Se emplearon técnicas de visualización de datos y análisis estadístico para facilitar la interpretación de los resultados y comunicar de manera efectiva las conclusiones a partes interesadas.* | |  |