Data Science: Entre la ciencia y el arte

Feature Engineering

En el campo del aprendizaje automático, la ingeniería de características juega un papel fundamental que a menudo marca la diferencia entre el éxito y el fracaso de un modelo predictivo. Andrew Ng, en su curso de aprendizaje automático en Stanford, ha destacado que la calidad de las características es lo que define el potencial de rendimiento de un modelo, mientras que los algoritmos simplemente se aproximan a ese potencial.

Sin embargo la ingeniería de características es una combinación de ciencia y arte. Implica la aplicación de técnicas estadísticas y matemáticas, pero también requiere creatividad y experimentación. Los científicos de datos dedican una cantidad significativa de tiempo a la preparación de datos, disfrutando del proceso de descubrir y crear características que puedan mejorar el rendimiento del modelo. Es una etapa esencial en el desarrollo de modelos predictivos efectivos. Dedicar tiempo a mejorar y seleccionar las características adecuadas puede ser la diferencia entre un modelo funcional y uno altamente competitivo.

¿Qué es la ingeniería de características?

La ingeniería de características es el proceso de transformar los datos brutos en características útiles que puedan ser interpretadas y utilizadas por los algoritmos de aprendizaje automático. Es el puente entre los datos crudos y los modelos, transformando la información en un formato que facilita la detección de patrones y la mejora del rendimiento del modelo. En esencia, características bien diseñadas pueden ofrecer una mejor representación de los problemas y mejorar la precisión de las predicciones en datos nuevos.

Las Cuatro Etapas Clave en la Ingeniería de Características

1. **Manejo de Valores Perdidos (missing):** Es crucial decidir cómo tratar los valores faltantes, ya sea a través de la imputación con valores medios, medianos o más frecuentes, o eliminando las instancias afectadas para evitar sesgos en el modelo.
2. **Detección y Manejo de Outliers:** Los valores extremos pueden distorsionar el modelo. Se pueden detectar y manejar usando técnicas como el método del rango intercuartil (IQR) para filtrar estos valores. Identificar y gestionar los valores extremos es esencial para evitar que distorsionen los resultados del modelo.
3. **Transformación de Características:**
    - **Escalado y normalización:** Escalar los datos para que todas las características tengan una magnitud similar es crucial, especialmente para algoritmos que dependen de distancias, como el k-NN. La normalización (escala 0-1) y normalización (media cero y varianza uno) son métodos comunes. Ten cuidado porque los términos escalar y normalizar son confusos en la literatura. Mira la formula mas que el nombre. Por ejemplo, el StandarScaler no escala, sino que normaliza (lo lleva a forma de campana de Gauss). Sí, normaliza (lleva todo a media cero y varianza uno). Min Max Scaler, escala.
    - **Transformaciones Logarítmicas:** Utilizadas para ajustar distribuciones sesgadas y hacerlas más normales, facilitando el análisis y la modelización.
    - La **discretización** de variables continuas: convierte variables continuas en características categóricas, lo que puede aumentar la robustez del modelo ante datos anómalos y facilitar la exploración de correlaciones.
    - **Transformación de Características Categóricas:** se pueden transformar utilizando métodos como la codificación de números naturales y la codificación one-hot

- **Transformación de Características Irregulares:** Además de las características numéricas y categóricas, a veces se encuentran características irregulares que pueden contener información valiosa sobre las muestras. Un ejemplo típico de esto es el número de identificación personal, como el número de cédula de identidad.

4. **Extracción de Características:**
    - **Características Estadísticas:** Generar características como medias o desviaciones estándar que resuman la información contenida en los datos originales.
    - **Combinación Cruzada de Características Categóricas:** permiten describir contenido más detallado y realizar un ajuste no lineal de características
    - **Correlaciones numéricas y combinaciones cruzadas:** Crear combinaciones de características mediante operaciones aritméticas para el caso de las numéricas. Se crean columnas adicionales combinando columnas entre sí.
    - **Características de Series Temporales:** Extraer patrones como tendencias y estacionalidades en datos temporales para mejorar la calidad de las predicciones.
    - **Características con Valores Múltiples:** Para manejar características con valores múltiples, se suelen emplear técnicas de **esparcimiento** o **vectorización**, que son comunes en el procesamiento de lenguaje natural.
5. **Selección de Características:**
    - **Importancia de las Características:** Utilizar modelos basados en árboles y otros métodos para determinar qué características son más relevantes para el modelo.
    - **Coeficientes en Modelos Lineales:** En modelos como la regresión logística, los coeficientes asociados a cada característica indican su influencia en las predicciones.

1. Preprocesamiento de Datos: Manejo de Valores Perdidos en Datos (Missing): ¿cómo manejarlos?

En el ámbito del aprendizaje automático, el tratamiento de valores perdidos es una tarea fundamental que se enfrenta tanto en competencias como en aplicaciones prácticas. Los valores faltantes pueden surgir por diversas razones, como fallos en la recopilación de datos, problemas técnicos o la negativa de los usuarios a proporcionar información. Aunque algoritmos como XGBoost y LightGBM tienen la capacidad de manejar valores faltantes directamente durante el entrenamiento, otros modelos como la regresión logística (LR), redes neuronales profundas (DNN), redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN) requieren una preparación previa de los datos para gestionar estos valores ausentes.

1.1 Identificación de Valores Perdidos

El primer paso en el manejo de valores faltantes es identificar las diversas formas en que estos pueden aparecer en el conjunto de datos. Aparte de los valores explícitos como None, NA o NaN, existen otros indicadores de datos faltantes, tales como -1 o -999. Además, algunas veces los datos ausentes pueden ser representados por valores que, aunque no presentes, tienen significados específicos:

- **Estado Civil:** Si un usuario no completa su estado civil, podría ser una señal de preocupación por la privacidad. En lugar de tratar estos datos como simplemente ausentes, se puede utilizar un valor específico como -1 para indicar que la información no fue proporcionada.
- **Experiencia de Conducción:** Si el campo sobre experiencia de conducción está vacío, puede interpretarse como que el usuario no posee un vehículo. En vez de dejar el campo en blanco, se puede rellenar con 0 para indicar ausencia de datos de manera que se ajuste al análisis.

1.2 Estrategias para Rellenar Valores Perdidos

Una vez identificados los valores perdidos y comprendida su naturaleza, es esencial rellenarlos adecuadamente para mantener la integridad de los datos y la calidad del modelo. Aquí se presentan algunas estrategias comunes:

Imputación Simple

Consiste en reemplazar los valores faltantes con la media, mediana o moda de la característica correspondiente. Aunque es una técnica sencilla y rápida, puede no captar la complejidad subyacente de los datos.

Imputación por Regresión

Utiliza técnicas de regresión para predecir los valores faltantes basándose en otras características del conjunto de datos. Este enfoque puede ser más preciso pero también más complejo.

Imputación por Vecinos Más Cercanos (KNN)

Rellena los valores faltantes utilizando los valores de las instancias más similares. Este método puede capturar patrones locales en los datos.

Categorías Especiales

En casos donde el valor faltante tiene un significado particular, se puede asignar una categoría especial para mantener la información implícita. Por ejemplo, en lugar de omitir los valores no proporcionados en un campo de estado civil, se puede usar un valor específico como -1.

1.3 Manejo de missings en variables Categóricas y Numéricas

Las estrategias para manejar los valores perdidos varían dependiendo de si se trata de características categóricas o numéricas:

Manejo de Características Categóricas

- **Relleno con un Nuevo Valor:** Se puede asignar un valor específico como 0, -1, o -∞ para indicar ausencia de datos, tratándolo como una categoría adicional durante el análisis.
- **Relleno con la Moda:** Reemplazar los valores faltantes con la moda (el valor más frecuente en el conjunto de datos). Esta técnica es útil si la distribución de categorías es sesgada y se desea conservar la estructura de datos.

Manejo de Características Numéricas

- **Promedio:** Rellenar los valores faltantes con la media de los valores conocidos es común, pero puede ser sensible a valores atípicos que distorsionan el análisis.
- **Mediana:** Usar la mediana para rellenar los valores perdidos es una alternativa robusta, especialmente útil en presencia de valores atípicos, ya que la mediana no se ve afectada por extremos.
- **Valores Estadísticos Alternativos:** Dependiendo del contexto, se pueden usar el valor máximo, mínimo o moda para rellenar datos faltantes. La elección debe basarse en un análisis específico del problema y la distribución de los datos.

1.4 Manejo de Datos Ordenados

Para datos ordenados, como las series temporales, se aplican técnicas específicas:

- **Interpolación:** Rellenar los valores faltantes utilizando los valores adyacentes anteriores o posteriores. Esta técnica es útil para mantener la continuidad en datos secuenciales y preservar patrones temporales.

1.5 Relleno Predictivo

Una técnica avanzada implica utilizar modelos predictivos para estimar los valores faltantes. En lugar de rellenar los valores perdidos con estadísticas simples, se puede construir un modelo predictivo que utilice otras características del conjunto de datos para estimar estos valores. Aunque este método es más complejo y computacionalmente intensivo, a menudo proporciona resultados más precisos y adaptados a la interacción entre variables.

En resumen, manejar valores perdidos adecuadamente es crucial para mantener la calidad del análisis y el rendimiento del modelo predictivo. La elección de la estrategia correcta depende del tipo de datos y del contexto específico, y puede requerir un equilibrio entre simplicidad y precisión.

2. Manejo de Valores atípicos (Outliers): ¿cómo manejarlos?

Los valores atípicos, o *outliers*, son puntos de datos que se desvían significativamente del comportamiento general del conjunto de datos. Estos valores pueden surgir por errores en la recolección de datos, variaciones naturales, o condiciones inusuales. Su presencia puede afectar negativamente la calidad de los modelos predictivos, por lo que es crucial identificarlos y decidir cómo manejarlos adecuadamente.

2.1 Identificación de Valores Atípicos

Para tratar con valores atípicos, el primer paso es identificarlos. Existen dos métodos comunes para este propósito:

Análisis Visual

Una de las formas más intuitivas de detectar valores atípicos es a través de la visualización de datos. Herramientas como gráficos de dispersión (scatter plots) permiten observar visualmente cómo se distribuyen los datos. Los puntos que se desvían considerablemente de las agrupaciones densas suelen ser considerados valores atípicos. Por ejemplo, en un gráfico de dispersión que muestra la relación entre dos variables, los puntos que están muy alejados de la tendencia general pueden ser identificados como atípicos.

Análisis Estadístico

Otra estrategia es el análisis estadístico, que utiliza métodos matemáticos para identificar anomalías en los datos:

- **Intervalo Intercuartílico (IQR):** Calcula el rango entre el primer cuartil (Q1) y el tercer cuartil (Q3). Los valores que están fuera del rango definido por 1.5 veces el IQR por encima de Q3 o por debajo de Q1 se consideran atípicos.
- **Desviación Estándar:** En el análisis de datos que siguen una distribución aproximadamente normal, los valores que se encuentran a más de 2 o 3 desviaciones estándar de la media pueden ser considerados atípicos.
- **Rango Extremo:** Utiliza los valores mínimo y máximo para detectar puntos que se encuentran en los extremos de la distribución.

2.2 Estrategias para Manejar Valores Atípicos

Una vez identificados, los valores atípicos deben ser tratados para evitar que distorsionen el análisis y el rendimiento del modelo. Aquí algunas estrategias comunes:

Eliminación de Registros con Valores Atípicos

Eliminar los registros que contienen valores atípicos puede ser una solución directa. Esto tiene la ventaja de eliminar la incertidumbre causada por estos datos extremos, pero también reduce el tamaño del conjunto de datos, lo que puede ser perjudicial si se pierde una cantidad significativa de información.

Considerar Valores Atípicos como Datos Faltantes

Otra estrategia es tratar los valores atípicos como datos faltantes y aplicar métodos de imputación. Esta aproximación centraliza los valores atípicos en una categoría y puede aumentar la disponibilidad de datos. Sin embargo, confundir los valores atípicos con datos faltantes puede afectar la precisión del análisis.

Corrección con Valores Promedio (Medianas)

Se pueden corregir los valores atípicos reemplazándolos con el valor promedio o la mediana de su categoría. Esto tiene ventajas similares a las de tratar los valores atípicos como datos faltantes, pero también comparte las mismas desventajas relacionadas con la posible pérdida de información específica.

No Procesar Valores Atípicos

En algunos casos, se puede optar por no procesar los valores atípicos y realizar el análisis directamente con el conjunto de datos que los incluye. La efectividad de este enfoque depende de la causa del valor atípico. Si es el resultado de un error en la entrada de datos, puede impactar negativamente en el modelo. Sin embargo, si el valor atípico refleja una situación real, no procesarlo puede permitir retener la información más auténtica y relevante.

3. Transformación de Características: ¿Qué hacer?

3.1 Escalar y normalizar variables continuas.

Escalar los datos para que todas las características tengan una magnitud similar es crucial, especialmente para algoritmos que dependen de distancias, como el k-NN. La normalización (escala 0-1) y normalización (media cero y varianza uno) son métodos comunes. Ten cuidado porque los términos escalar y normalizar son confusos en la literatura. Mira la formula mas que el nombre. Por ejemplo, el StandarScaler no escala, sino que normaliza (lo lleva a forma de campana de Gauss). Sí, normaliza (lleva todo a media cero y varianza uno). Min Max Scaler, escala.

La **no-dimensionalización** se refiere a la conversión de datos con diferentes especificaciones a una especificación común. Entre los métodos comunes de no-dimensionalización se incluyen **escalado** y **normalización**.

- **Escalado:** Este método utiliza la información de los valores extremos para ajustar el intervalo de los valores de las características a un rango específico, como [0, 1].
- **Normalización:** Este método se basa en la suposición de que los valores de las características siguen una distribución normal. La normalización transforma los valores para que se ajusten a una distribución normal estándar, con media 0 y desviación estándar 1.

  La transformación de características es crucial para construir ciertos modelos, como la regresión lineal, KNN y redes neuronales. Estos modelos se benefician de una escala uniforme entre las características para funcionar correctamente. En contraste, los modelos basados en árboles de decisión (como Random Forest y Gradient Boosting) no requieren esta transformación, lo que contribuye a su popularidad.

Por ejemplo, Supongamos que queremos predecir el costo de apartamentos basándonos en dos variables: el número de habitaciones y la distancia al centro de la ciudad. El número de habitaciones rara vez supera 5, mientras que la distancia al centro puede variar varios kilómetros. Sin normalización, la distancia al centro, con su rango mucho mayor, podría dominar la predicción, haciendo que la variable del número de habitaciones sea ineficaz. Por lo tanto, es necesario **escalar** y **normalizar** estas variables para que estén en la misma escala y asegurar que ambas características contribuyan de manera equitativa al modelo.

3.2 Transformación Logarítmica

La **transformación logarítmica** es una técnica utilizada para corregir la asimetría en los datos, acercándolos a una distribución normal. Esto es especialmente útil porque muchos modelos de aprendizaje automático tienen dificultades para manejar datos con distribuciones no normales, como los datos sesgados a la derecha.

El propósito de agregar 1 es evitar problemas con valores de datos que sean cero, asegurando que xxx sea siempre positivo. Aunque tomar logaritmos no cambia la naturaleza ni la correlación de los datos, sí comprime la escala de las variables, lo que puede hacer que los datos sean más estables y reducir problemas como la colinealidad y la heterocedasticidad en el modelo.

3.3 La Transformación Box-Cox: Esta es una técnica avanzada que busca automáticamente la mejor función de transformación para aproximar una distribución normal. Aunque no es comúnmente utilizada en competiciones, puede ser de interés para quienes desean explorar métodos más sofisticados.

3.4 Discretización de Variables Continuas

La **discretización** convierte variables continuas en características categóricas, lo que puede aumentar la robustez del modelo ante datos anómalos y facilitar la exploración de correlaciones. Por ejemplo, al discretizar la característica de edad: si la edad es mayor de 30, se asigna el valor 1; de lo contrario, se asigna 0. Sin discretización, un valor anómalo como "edad 300 años" podría afectar significativamente al modelo. La discretización también permite la combinación cruzada de características.

Métodos de Discretización:

- **Discretización No Supervisada:** Esta técnica divide las variables continuas en intervalos, reduciendo el impacto del ruido. Los métodos más comunes son:
    - **Equifrecuencia:** Se seleccionan los valores de frontera de manera que cada intervalo contenga aproximadamente el mismo número de instancias. Por ejemplo, si se divide en 10 intervalos, cada uno debe contener alrededor del 10% de las instancias. Este método puede transformar los datos en una distribución uniforme.
    - **Isometría:** Divide las instancias desde el mínimo hasta el máximo en N partes iguales, con el espacio entre cada parte siendo igual. Solo se considera la frontera, por lo que el número de instancias por parte puede variar. La isometría conserva mejor la distribución original de los datos y, con más intervalos, se mantiene más la apariencia original de los datos.
- **Discretización Supervisada:** Este método es eficaz para distinguir entre diferentes objetivos. Se usa comúnmente para discretizar utilizando modelos de árboles, como en el modelo clásico GBDT + LR mostrado en la figura. En este enfoque, primero se utiliza GBDT para convertir valores continuos en valores discretos. Se entrena un modelo de LightGBM con todas las instancias continuas y las etiquetas de salida del conjunto de entrenamiento, creando árboles de decisión con nodos hoja. Los valores discretos se obtienen a partir de la posición en los nodos hoja, generando características discretas basadas en la ubicación de cada muestra en el árbol.

3.5 Transformación de Características Categóricas

En el análisis de datos, las características no siempre son numéricas; también pueden ser categóricas. Estas características categóricas se pueden transformar utilizando métodos como la codificación de números naturales y la codificación one-hot.

Codificación de Números Naturales

La **codificación de números naturales** se aplica a características categóricas que tienen un orden intrínseco. Este método convierte cada categoría en un número entero, preservando el orden secuencial. Por ejemplo, en una característica que clasifica niveles educativos como Primaria, Secundaria y Terciaria, se puede asignar un número entero a cada nivel para reflejar su orden.

Este tipo de codificación permite un bajo consumo de memoria y reduce el tiempo de entrenamiento del modelo, ya que el uso de números enteros en lugar de múltiples columnas puede hacer el procesamiento más eficiente. Sin embargo, si el modelo no maneja correctamente el orden numérico, podría haber pérdida de información sobre las relaciones entre categorías.

Métodos para Implementar Codificación de Números Naturales:

LabelEncoder

La función LabelEncoder convierte las categorías en números enteros. **También** Se pueden asignar manualmente números a las categorías usando técnicas personalizadas.

Codificación One-Hot

La **codificación one-hot** es adecuada para características categóricas sin un orden intrínseco. En lugar de representar las categorías como números enteros, este método convierte cada categoría en una columna binaria separada, donde cada columna indica la presencia o ausencia de la categoría mediante valores 1 o 0. Por ejemplo, para una característica categórica que incluye los colores rojo, azul y verde, se crean tres columnas binarias, una para cada color. La clase OneHotEncoder facilita la conversión de características categóricas en una representación binaria.

Este método preserva toda la información de las categorías y es especialmente útil para características donde no existe un orden lógico entre las opciones. Aunque puede generar un gran número de columnas, lo que podría aumentar el consumo de memoria y el tiempo de procesamiento, es ideal para mantener la integridad de las características categóricas.

3.6 Transformación de Características Irregulares

Además de las características numéricas y categóricas, a veces se encuentran características irregulares que pueden contener información valiosa sobre las muestras. Un ejemplo típico de esto es el número de identificación personal, como el número de cédula de identidad.

Ejemplo de Número de Identificación

Tomando como ejemplo el número de identificación nacional de China, que está regido por el estándar nacional GB 11643-1999, este número es una combinación de códigos que proporciona información estructurada. El número de identificación consta de:

- Un código de dirección de seis dígitos.
- Un código de fecha de nacimiento de ocho dígitos.
- Un código secuencial de tres dígitos.
- Un código de verificación de un dígito.

Dentro del código secuencial, los números impares suelen estar asignados a hombres, mientras que los números pares se asignan a mujeres. El código de verificación se calcula utilizando el método de verificación ISO 7064:1983, MOD 11-2, basado en los 17 dígitos anteriores.

Extracción de Información

Aunque el número de identificación contiene información confidencial, en teoría, se puede extraer una serie de datos útiles:

- **Lugar de Nacimiento:** A partir del código de dirección.
- **Edad:** Basado en el código de fecha de nacimiento.
- **Género:** Determinado por el código secuencial.

4. Extracción de Características ¿Cómo hacerlo?

Los modelos de aprendizaje automático a menudo tienen dificultades para identificar patrones complejos, especialmente cuando estos patrones implican interacciones entre diferentes combinaciones de características. Para ayudar a los modelos a aprender de manera más efectiva, podemos crear características adicionales basadas en un análisis intuitivo del conjunto de datos y en la comprensión del negocio. A continuación, se presentan métodos para la extracción de características en datos estructurados. (Los datos estructurados tienen tipos de datos claramente definidos, mientras que los datos no estructurados incluyen datos difíciles de buscar, como audio, video y fotos.)

4.1 **Características Estadísticas Relacionadas con Categorías**

Las características categóricas, también conocidas como características discretas, no solo tienen un significado específico para cada atributo de categoría, sino que también pueden generar características estadísticas continuas para extraer información más valiosa. Entre las técnicas comunes se encuentran la codificación de objetivos, el conteo, el número único (nunique) y el cálculo de proporciones. También es posible construir características más detalladas mediante combinaciones cruzadas entre características categóricas.

Codificación de Objetivos

La codificación de objetivos implica codificar las características categóricas utilizando estadísticas de las variables objetivo (etiquetas). Esto se traduce en una construcción de características supervisada basada en las variables objetivo. En el caso de problemas de clasificación, se puede contar el número de muestras positivas, el número de muestras negativas o la proporción entre estas; en problemas de regresión, se pueden contar la media, la mediana y los valores extremos del objetivo. La codificación de objetivos puede ser una buena alternativa a las características categóricas originales o puede ser utilizada como una nueva característica.

Es crucial evitar que la codificación de objetivos revele información sobre el conjunto de verificación. Todas las características basadas en la codificación de objetivos deben calcularse utilizando únicamente el conjunto de entrenamiento. Para minimizar el riesgo de filtración de información, se puede utilizar el método de estadística de K-fold cross-validation. Por ejemplo, si se divide la muestra en cinco partes, para cada parte de los datos se utilizan las otras cuatro partes para calcular la frecuencia, proporción o media del objetivo correspondiente a cada valor de categoría. En otras palabras, los datos desconocidos (una parte) utilizan características de los datos conocidos (cuatro partes).

Los métodos de codificación de objetivos suelen ser efectivos para características categóricas con baja cardinalidad, pero para características con alta cardinalidad, existe el riesgo de sobreajuste. Esto se debe a que algunas categorías con baja frecuencia pueden no proporcionar resultados representativos. Generalmente, se puede añadir suavizado para reducir el riesgo de sobreajuste. Si se maneja adecuadamente, la codificación de objetivos es un método de codificación y construcción de características muy eficaz tanto para modelos lineales como no lineales.

4.2 Características Estadísticas Relacionadas con Categorías

Las características categóricas, también conocidas como características discretas, pueden generar información valiosa a través de estadísticas continuas. Entre las técnicas comunes se encuentran el conteo, el número único (nunique) y la proporción. Estas técnicas pueden revelar patrones y relaciones importantes en los datos.

- *Conteo (Count):* El conteo se utiliza para calcular la frecuencia de ocurrencia de una característica categórica. Este método proporciona una medida directa de cuántas veces aparece cada categoría en el conjunto de datos.
- *Número Único (Nunique):* Esta técnica mide la diversidad dentro de una categoría. Por ejemplo, en el contexto de una predicción de tasa de clics en anuncios, el número único puede indicar cuántos tipos diferentes de anuncios ha visto un usuario, lo cual refleja el rango de interés del usuario.
- *Proporción (Ratio):* La proporción calcula la relación entre la frecuencia de una categoría y la frecuencia total. Por ejemplo, la proporción puede mostrar el nivel de preferencia de un usuario por un tipo específico de anuncio, dividiendo la frecuencia de clics en ese tipo de anuncio por la frecuencia total de clics del usuario en todos los anuncios.

Combinación Cruzada de Características Categóricas

Las combinaciones cruzadas de características categóricas permiten describir contenido más detallado y realizar un ajuste no lineal de características. Combinar características categóricas es una tarea crucial en muchas competencias de datos. Por ejemplo, combinar la edad del usuario y el género del usuario en una nueva característica como "edad_genero" puede revelar patrones más complejos.

Se pueden combinar dos o tres características categóricas, conocidas como combinaciones de segundo o tercer orden. Esto implica realizar un producto cartesiano de las características categóricas para generar nuevas características categóricas. Si se tienen 10 características categóricas y se consideran todas las combinaciones de segundo orden, se pueden producir 45 combinaciones diferentes.

No todas las combinaciones deben ser consideradas. Se deben analizar dos aspectos:

- **Lógica del Negocio:** Algunas combinaciones pueden no tener sentido en el contexto específico del problema. Por ejemplo, combinar la versión del sistema operativo del usuario con la ciudad del usuario puede no proporcionar información útil.
- **Cardinalidad de las Características:** Si la cardinalidad es demasiado alta, es posible que muchas categorías aparezcan solo una vez. En este caso, cada categoría se entrenará solo una vez, lo que resultará en un bajo nivel de confianza en el peso correspondiente a esa característica.

4.3 Correlación Numérica de Características Estadísticas

Las características numéricas, como el precio de vivienda, el volumen de ventas, el número de clics, el número de comentarios y la temperatura, son continuas y, a diferencia de las características categóricas, sus valores son significativos y generalmente pueden ser introducidos directamente en el modelo para el entrenamiento sin necesidad de procesamiento adicional. Aparte de las transformaciones numéricas previas, existen varios métodos comunes para construir características numéricas adicionales.

- **Combinación Cruzada Entre Características Numéricas:** A diferencia de las combinaciones cruzadas entre características categóricas, las combinaciones cruzadas de características numéricas se realizan mediante operaciones aritméticas como suma, resta, multiplicación y división. Es esencial combinar la comprensión del negocio y el análisis de datos para construir estas características de manera reflexiva. Por ejemplo, si se conocen el tamaño de la vivienda (en metros cuadrados) y el precio de venta, se puede construir una característica que calcule el precio promedio por metro cuadrado. De manera similar, si se dispone del monto de consumo mensual del usuario durante los últimos tres meses, se pueden calcular el monto total de consumo y el promedio de consumo para reflejar la capacidad general de gasto del usuario.
- **Combinación Cruzada Entre Características Categóricas y Numéricas:** Además de combinar características categóricas entre sí y características numéricas entre sí, también es posible combinar características categóricas con características numéricas. Estas características suelen calcular estadísticas de características numéricas dentro de categorías específicas, como la media, mediana y valores extremos. Por ejemplo, se puede calcular la media del consumo de electricidad para cada tipo de usuario en un conjunto de datos categorizado por tipo de usuario.
- **Estadísticas por Fila de Características Relacionadas:** Este enfoque es similar al cruce de características, ya que combina información de múltiples columnas. Sin embargo, las estadísticas por fila pueden incluir más columnas cuando se construyen. Este método consiste en contar múltiples columnas por fila, como el número de valores 0, valores nulos, valores positivos y negativos, o calcular medias, medianas, valores extremos, o sumas. Por ejemplo, para un conjunto de datos con características relacionadas como el monto de consumo y el consumo de electricidad mensual, se pueden generar estadísticas que reflejen el comportamiento general del consumo. En datos industriales, esto puede incluir la temperatura y concentración en cada etapa de experimentos químicos, permitiendo extraer características valiosas del análisis de los cambios en los valores de múltiples columnas.

4.4 Características de Tiempo (Series Temporales)

- En datos reales, la característica de tiempo generalmente se proporciona como un atributo de marca temporal (timestamp), el cual suele necesitar ser descompuesto en múltiples dimensiones para un análisis más detallado. Estas dimensiones incluyen el año, mes, día, hora, minuto y segundo. Además, si los datos provienen de diferentes zonas geográficas, es crucial estandarizar los datos utilizando las zonas horarias correspondientes.
- Además de descomponer la marca temporal en características básicas, se pueden construir características adicionales basadas en la diferencia de tiempo. Esto implica calcular la diferencia numérica entre el momento de cada muestra y un momento futuro o de referencia. Esta diferencia convierte la característica de tiempo en un valor continuo. Ejemplos de esto incluyen calcular la diferencia de tiempo entre la fecha del primer comportamiento de un usuario y la fecha de registro del usuario, así como la diferencia de tiempo entre el comportamiento actual del usuario y su último comportamiento.

4.5 Características con Valores Múltiples

En competiciones reales, es común encontrarse con situaciones en las que una fila en una columna de características contiene múltiples atributos, lo que se denomina característica con valores múltiples. Un ejemplo es la categoría de intereses en la Competencia de Algoritmos de Publicidad de Tencent 2018, donde cada grupo de características de interés contiene varios identificadores de interés. Para manejar características con valores múltiples, se suelen emplear técnicas de **esparcimiento** o **vectorización**, que son comunes en el procesamiento de lenguaje natural.

Una de las maneras más básicas de tratar con características con valores múltiples es **expandir completamente** estas características. Esto implica convertir los atributos contenidos en la lista de características en una matriz dispersa de dimensiones correspondientes al número de atributos. Por ejemplo, usando la función CountVectorizer de sklearn, se puede expandir fácilmente características con valores múltiples, enfocándose únicamente en la frecuencia de cada atributo en la característica.

En otro escenario, como en la Competencia de Algoritmos de Publicidad de Tencent 2020, donde se requiere predecir las pestañas de atributos de un usuario en función del historial de clics del usuario, la secuencia de clics del usuario es crucial. Para construir la secuencia de clics histórica del usuario, además de usar métodos como TF-IDF, se pueden extraer representaciones incrustadas de productos o anuncios en la secuencia de clics. Técnicas como Word2Vec y DeepWalk pueden obtener la representación vectorial incrustada. Una vez obtenidos los vectores incrustados, se pueden agregar estadísticas sobre estos vectores en la secuencia, tratando cada clic como igualmente importante. Este enfoque es relativamente rudimentario; se pueden introducir factores de atenuación temporal o usar modelos secuenciales como RNN, LSTM, GRU, y aplicar métodos de procesamiento de lenguaje natural (NLP) para un tratamiento más avanzado.

Hasta aquí, se han presentado métodos básicos para construir características. Existen muchas otras técnicas no mencionadas, como características espaciales, series temporales, características textuales, así como métodos de clustering y reducción de dimensionalidad, que se tratarán en detalle en los capítulos siguientes al abordar problemas específicos.

5. Selección de Características ¿Cómo hacerlo?

Cuando se añaden nuevas características, es esencial verificar si realmente mejoran la precisión de la predicción del modelo. Esto es necesario para evitar la inclusión de características inútiles que solo aumentarían la complejidad del algoritmo. La selección de características es el proceso mediante el cual se identifican y eliminan características innecesarias, irrelevantes o redundantes para mejorar el rendimiento del modelo. Los métodos de selección de características incluyen análisis de correlación previa y análisis de importancia posterior.

5.1 Análisis de Correlación de Características

El análisis de correlación de características utiliza estadísticas para medir la correlación entre características. Las características se ordenan según sus puntuaciones para decidir si se mantienen o eliminan. Los métodos comunes de análisis de correlación incluyen:

- **Coeficiente de correlación de Pearson**: Mide la correlación lineal entre variables y ayuda a solucionar el problema de las variables colineales. Las variables colineales, que tienen alta correlación entre sí, pueden reducir la capacidad de aprendizaje, interpretabilidad y rendimiento general del modelo. La eliminación de variables colineales es una práctica valiosa.
- **Prueba de Chi-cuadrado**: Evalúa la independencia entre una variable característica y la variable dependiente. Es útil para problemas de clasificación y puede indicar si una característica es independiente de la etiqueta, en cuyo caso puede ser eliminada.
- **Método de información mutua**: Mide la relación entre dos variables en una distribución conjunta. Se basa en la divergencia KL y la ganancia de información. La información mutua puede ser difícil de calcular para variables continuas y es sensible a la discretización.

5.2 Análisis de Importancia de Características

La importancia de las características a menudo se evalúa utilizando modelos de árboles, como XGBoost, que puede devolver la importancia de las características mediante tres métodos principales:

- **Método de peso**: Calcula la frecuencia con la que una característica es utilizada como divisor en todos los árboles del modelo. La importancia se basa en el número de veces que la característica es seleccionada.
- **Método de ganancia**: Representa el promedio de la ganancia. Calcula la suma de las ganancias de información de una característica como nodo de división en todos los árboles, dividida por la frecuencia de aparición de la característica.
- **Método de cobertura**: Mide la tasa de cobertura de la característica en cada árbol, considerando la suma de las segundas derivadas de las muestras donde la característica se asigna al nodo, y el estándar de medición es el valor promedio de cobertura.

Uso de Habilidades

Aunque la importancia de las características ayuda a analizar rápidamente su valor en el proceso de entrenamiento del modelo, no debe ser utilizada como una referencia absoluta. En general, se pueden seleccionar características con alta importancia para el análisis y expansión, mientras que las características con baja importancia pueden considerarse para eliminación, observando su efecto en el rendimiento y realizando juicios adicionales.

5.3 Métodos de Encapsulación

Los métodos de encapsulación son técnicas de selección de características que consideran el proceso de selección como un problema de búsqueda. En este enfoque, diferentes combinaciones de características se preparan, evalúan y comparan para encontrar el

subconjunto óptimo de características. Los métodos de encapsulación pueden ser sistemáticos, aleatorios, o heurísticos. A continuación, se describen dos métodos de encapsulación comúnmente utilizados:

- **Método Heurístico**: Este método se divide en dos tipos principales: búsqueda hacia adelante (forward search) y búsqueda hacia atrás (backward search).
    - **Búsqueda Hacia Adelante**: En este enfoque, se selecciona incrementalmente una de las características no seleccionadas y se añade al conjunto de características. Se repite este proceso hasta que el número de características en el conjunto alcanza un umbral inicial. La búsqueda busca seleccionar el subconjunto de características que minimiza la tasa de error.
    - **Búsqueda Hacia Atrás**: Este método comienza con el conjunto completo de características y elimina una característica a la vez, evaluando el rendimiento en cada paso. Se continúa hasta que se alcanza el umbral inicial de características, seleccionando el subconjunto que ofrece el mejor rendimiento.
    - **Mejoras a la Búsqueda Heurística**: Para superar la limitación de la optimización local en los métodos heurísticos, se puede aplicar un método de recocido simulado (simulated annealing). Este método no descarta características nuevas si no mejoran el rendimiento inmediatamente, sino que ajusta sus pesos y las incluye en el conjunto de características seleccionadas.

  Los métodos heurísticos son comunes en competiciones debido a su capacidad para encontrar soluciones razonables, aunque son intensivos en tiempo y recursos. Generalmente se utilizan cuando el rendimiento en línea y fuera de línea es comparable y el tamaño del conjunto de datos no es muy grande.

- **Método de Eliminación Recursiva de Características (RFE)**: Este método utiliza un modelo base para realizar múltiples rondas de entrenamiento. En cada ronda, se eliminan las características con los coeficientes de peso más bajos y se entrena de nuevo con el conjunto de características restante. El proceso se repite hasta que se obtiene el número deseado de características. La biblioteca feature_selection en Python ofrece la clase RFE para implementar este método.

Estos métodos permiten seleccionar subconjuntos de características que mejoran el rendimiento del modelo al centrarse en las características más relevantes y eliminar las menos útiles.

Uso de Métodos de Encapsulación

Cuando se utilizan métodos de encapsulación para la selección de características, no siempre es aconsejable entrenar con el conjunto de datos completo, especialmente en el caso de grandes volúmenes de datos. Es más prudente tomar una muestra representativa del conjunto de datos y aplicar los métodos de encapsulación en esta muestra más pequeña. Esto no solo optimiza el tiempo y los recursos, sino que también puede facilitar la evaluación de diferentes combinaciones de características de manera más eficiente.

Recomendaciones para la Selección de Características

1. **Priorizar Métodos de Importancia de Características**:
    - **Métodos Basados en la Importancia**: Los métodos que evalúan la importancia de las características (como los métodos basados en modelos de árboles) son generalmente preferidos, ya que proporcionan una medida directa de la utilidad de cada característica en la mejora del rendimiento del modelo.
    - **Métodos de Correlación de Características**: Estos pueden ser útiles, pero a menudo deben ser secundarios a la evaluación de la importancia. Esto se debe a que los métodos de correlación pueden no capturar las interacciones complejas entre características y el objetivo.
2. **Manejo de Características Falsamente Asociadas**:
    - **Método de Importancia Nula**: Este es un enfoque clásico utilizado para evaluar la validez de la importancia de las características. La idea básica es comparar la importancia de características construidas con etiquetas correctas y etiquetas alteradas.
        - **Proceso del Método de Importancia Nula**:
            1. **Entrenamiento Inicial**: Se entrena un modelo de árbol (o cualquier otro modelo adecuado) usando las características originales y las etiquetas correctas para obtener una puntuación de importancia de características.
            2. **Desviación de Etiquetas**: Se altera aleatoriamente las etiquetas mientras se mantienen las características originales y se entrena el modelo nuevamente para obtener una segunda puntuación de importancia de características.
            3. **Comparación de Puntuaciones**: Se comparan las dos puntuaciones de importancia. Si la puntuación obtenida con las etiquetas alteradas es mayor o igual que la obtenida con las etiquetas correctas, esto sugiere que la característica no aporta valor real y puede considerarse inútil.