

Texto Sustentatorio para el Modelo de Predicción de Horas para Realizar un Tatuaje

Creamos algunos objetivos a alcanzar con este proyecto:

Objetivo general:

Analizar los datos del estudio de tatuajes y usar un modelo de predicción para estimar cuánto puede durar una sesión según las características del cliente y del tatuaje, y mostrar todo de forma clara en un dashboard de Power BI.

Específicos:

- Entender qué variables influyen más en la duración de una sesión de tatuaje.
- Usar el modelo entrenado para generar una columna con la duración estimada y trabajar esa información dentro del dashboard.
- Crear visualizaciones que permitan comparar estilos, artistas, tamaños y tipos de sesión para ver si realmente afectan el tiempo.
- Mostrar de forma sencilla y visual si la predicción del modelo tiene coherencia con los datos reales.
- Organizar la información de manera que cualquier persona pueda saber qué tipo de tatuaje puede tomar más tiempo y en qué condiciones.

El presente trabajo tiene como objetivo la creación de un modelo predictivo para estimar el tiempo que se tardará en realizar un tatuaje, basándose en diversas características que pueden influir en este proceso. La efectividad del modelo alcanzó un 86%, lo que indica un buen desempeño en la predicción de las horas necesarias, aunque siempre es importante realizar ajustes y mejoras continuas en el modelo a medida que se disponen de más datos, únicamente tiene falencias al predecir el tiempo medio, ya que lo confunde un poco con el tiempo corto, pero el corto y el largo funcionan casi perfecto.

El modelo fue entrenado utilizando una base de datos con 2,000 registros. Esta base de datos original contiene información clave, como el tamaño del tatuaje, el estilo, el género, la edad, 4 tatuadores temporales y otros factores que podrían afectar el tiempo requerido para realizar un tatuaje. Sin embargo, para poder realizar pruebas adicionales y expandir el análisis, se decidió crear una base de datos imaginaria con las mismas características y tamaño (2,000 registros).

La creación de esta base de datos ficticia fue fundamental para garantizar que el modelo pudiese generalizar su capacidad predictiva a un conjunto de datos adicional y verificar que el rendimiento no estuviera sesgado por las características específicas de la base original. Para mantener la coherencia, se replicaron las mismas distribuciones y patrones de las

variables presentes en los datos reales. Esto permitió validar el modelo de manera robusta y asegurar que las predicciones fueran confiables.

Con el modelo entrenado y probado, se procedió a generar un archivo de formato .pkl (pickle) que contiene el modelo entrenado. Este archivo se creó para integrarlo posteriormente en un nuevo entorno de trabajo .ipynb. El nuevo notebook y archico pikle en Python permite cargar el modelo y realizar predicciones sobre la nueva base de datos imaginaria. A través de este archivo .pkl, se puede acceder al modelo sin necesidad de reentrenarlo, lo que optimiza el tiempo de ejecución y facilita su implementación en otras aplicaciones.

Además, se preparó la información para ser visualizada en Power BI, permitiendo crear un panel interactivo que facilita la interpretación de los resultados y las predicciones del modelo. Este informe de Power BI proporciona a los usuarios una visión clara y dinámica de las estimaciones del tiempo de tatuaje, teniendo en cuenta diferentes variables y características del diseño. Las visualizaciones incluyen gráficos de barras, distribuciones de tiempo estimado y comparaciones entre las horas predichas y las horas reales, lo cual es clave para evaluar la precisión y el rendimiento del modelo.

Para este proyecto primero intentamos predecir otras cosas, como la satisfacción del cliente o el final rate, pero ese camino no funcionó nada bien. La satisfacción era casi imposible de estimar con las variables que teníamos, porque realmente no había una relación directa entre cosas como el estilo del tatuaje o el día de la semana y cómo queda de feliz la persona al final. Todo era súper disperso y el modelo no encontraba patrones reales.

Después probamos con el final rate, pero ahí el problema era que el dato era numérico y demasiado variable. El modelo no lograba aprender algo estable ni consistente, así que tampoco era una opción viable.

Por eso terminamos enfocándonos en algo mucho más lógico: predecir la duración estimada de la sesión. Esa variable sí tiene una relación clara con cosas como el tamaño del tatuaje, el estilo, el artista, el tipo de cita, etc. Aquí sí había patrones fuertes y repetitivos.

Ya con eso definido, usamos un modelo Random Forest, principalmente porque: Funciona muy bien con datos mixtos, numéricos y categóricos, tolera bien el ruido y diferencias dentro de las categorías, no requiere tantos ajustes complicados para empezar a dar buenos resultados, fue el que mostró mejor rendimiento en comparación con otras pruebas.

Intentamos otras alternativas (como regresiones o modelos más simples), pero no tenían buen desempeño porque la variable objetivo es categórica hablando de tiempo corto>=3 horas, medio>=6 horas, largo más de 6 horas, y los Random Forest clasifican muy bien cuando la relación entre variables es más compleja.

Así que este fue el único modelo que realmente se comportó bien, aprendió patrones útiles y nos permitió generar la columna de predicción que luego usamos en Power BI para todas las visualizaciones.