SEGUNDA ENTREGA DE PROYECTO IA

Presentado por:

Karol Melissa Reyes Anaya

Jorge Andrés Cardeño Devia

Profesor

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín 2023-2

Home Credit Default Risk

Enlace de la competición en kaggle:

https://www.kaggle.com/competitions/home-credit-default-risk/overview

El proyecto Home Credit Default Risk tiene como objetivo principal anticipar la probabilidad de que los solicitantes de crédito incumplan sus compromisos de pago. Para lograr este propósito, resulta de vital importancia llevar a cabo una cuidadosa manipulación de los datos antes de someterlos a modelado y análisis. En este informe, se describen avances significativos en las técnicas de manipulación de datos que pueden influir de manera positiva en la calidad de las predicciones.

En la fase inicial de la manipulación de datos, se procede a la carga de los archivos de datos de entrenamiento y prueba. Esto se realiza mediante el uso de la biblioteca Pandas de Python y la función "read_csv". A continuación, se fusionan estos conjuntos de datos en un único DataFrame con la función "concat", lo que facilita una manipulación más coherente y eficaz de los datos.

Una vez que los datos se han combinado, se inicia una serie de manipulaciones básicas que, en este avance particular, se centran en la generación de nuevas variables a partir de las ya existentes. Se destacan dos variables novedosas: "income_per_person" y "credit_term". La primera de estas variables se calcula mediante la división del ingreso total entre el número de miembros en la familia, mientras que la segunda variable resulta de la división del monto del préstamo entre el plazo en días. La incorporación de estas nuevas variables promete enriquecer las predicciones relacionadas con el incumplimiento de pago. La variable "income_per_person" podría resultar un indicador valioso del riesgo de impago, ya que los hogares con ingresos más bajos podrían enfrentar mayores dificultades para cumplir con sus compromisos financieros. Además, la variable "credit_term" podría ser útil para identificar a los solicitantes que solicitan plazos de pago poco realistas, lo que a su vez podría indicar un riesgo más elevado de incumplimiento.

La creación de estas nuevas variables debe llevarse a cabo con gran cuidado y consideración, basándose en una sólida lógica y conocimiento del dominio en cuestión. Además, es imperativo asegurarse de que estas nuevas variables sean realmente relevantes y aporten valor al modelo de predicción.

El proceso de manipulación de datos es la gestión de los valores faltantes. La presencia de datos faltantes es un fenómeno común en cualquier conjunto de datos y puede causar problemas en los modelos de predicción si no se aborda de manera adecuada. En este proyecto en particular, se empleó una técnica de imputación que se basa en la media y la mediana para rellenar los valores faltantes en el conjunto de datos. Esta estrategia, aunque sencilla, ha demostrado ser eficaz.

Básicamente, consiste en sustituir los valores faltantes por la media o la mediana de los valores existentes en la misma columna. Esto garantiza que los datos estén completos y evita la introducción de sesgos o errores en el modelo de predicción.

Es importante tener en cuenta que la técnica de imputación por media y mediana no es necesariamente la más adecuada para todos los conjuntos de datos. En algunas situaciones, puede requerirse el uso de enfoques más avanzados, como la imputación basada en modelos o la imputación múltiple.

El proceso de manipulación de datos desempeña un papel fundamental en la creación de modelos de predicción de riesgo crediticio, como el proyecto Home Credit Default Risk. Este informe ha puesto de relieve varios avances en la manipulación de datos, que incluyen la creación de nuevas variables, la imputación de valores faltantes y la eliminación de variables irrelevantes. Estos avances persiguen el objetivo de mejorar la calidad de las predicciones y, por lo tanto, contribuir a la identificación temprana de solicitantes de crédito de alto riesgo.

En conclusión, los avances en la manipulación de datos presentados en este informe desempeñan un papel crítico en el éxito del proyecto Home Credit Default Risk. Al mejorar la calidad de las predicciones, se espera que estos avances contribuyan a la identificación temprana de solicitantes de crédito de alto riesgo. Esto, a su vez, puede ayudar a reducir el riesgo de incumplimiento de créditos y mejorar la rentabilidad de la compañía.