SEGUNDA ENTREGA DE PROYECTO IA

Presentado por:

Sebastian Castro Bolaños

Profesor

Raúl Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

Medellín 2023-2

RIESGO DE INCUMPLIMIENTO DE CRÉDITO HIPOTECARIO

El proyecto elegido para el semestre en curso tiene como objetivo predecir la probabilidad de impago de los solicitantes de un crédito. Para llegar a esto, es fundamental realizar una manipulación adecuada de los datos antes de su modelado y análisis, pues la correcta manipulación de los datos es de vital importancia en el entrenamiento de máquina. En este informe, se describe un avance en el filtrado de datos que puede ayudar a mejorar la calidad de las predicciones, además se tiene pensado la creación de nuevas variables a partir de la lógica del problema.

El primer paso en la manipulación de datos es cargar los archivos de datos proporcionados por la competencia. Esto se hace a partir de 'gdown', que nos permite descargar archivos almacenados en el drive. Posterior a esto se descomprime el zip y se procede a analizar los diferentes datos a partir de la biblioteca Pandas y la función 'read_csv'. Donde, por consideraciones de importancia de datos se determina trabajar con el data set 'application_train', que es el que contiene la columna Target, que es la que nos piden predecir en la competencia.

En este avance se ha realizado el filtrado de información, pues al tener un data set tan grande (307511,121), es importante analizar qué información proporcionada es relevante para el ejercicio a desarrollar.

Es por esto que se realiza diferentes métodos de filtrado de columnas y filas que se puede evidenciar más a detalle en el Notebook, llegando finalmente a un total de 307511 filas y 51 columnas, siendo lo más rigurosos posibles para así no alterar la calidad de la futura predicción.

Una vez filtrado los datos se procede a rellenar los datos faltantes con la media y la moda de cada columna. La técnica de imputación por media y mediana es un método simple pero efectivo para tratar los datos faltantes. En esencia, implica reemplazar los valores faltantes con la media o mediana de los valores existentes en la misma columna. Esto ayuda a garantizar que los datos estén completos y que no se introduzcan sesgos o errores en el modelo de predicción. Es importante tener en cuenta que la técnica de imputación por media y mediana puede no ser adecuada para todos los conjuntos de datos. En algunos casos, puede ser necesario utilizar técnicas más avanzadas de imputación, como la imputación por modelo o la imputación múltiple.

De esta manera, al tener el mismo número de filas para cada columna, se procede a crear nuevas variables de entrenamiento a partir de las variables existentes, esto con la finalidad de mejorar la calidad de las predicciones de incumplimiento de pago.

Es importante tener en cuenta que la creación de nuevas variables debe realizarse con cuidado y en función de la lógica y el conocimiento experto del dominio. Además, es necesario asegurarse de que estas nuevas variables sean relevantes y útiles para el modelo de predicción.

En conclusión, el proceso de manipulación de datos es esencial en la creación de modelos de predicción de riesgo crediticio. En este informe, se presentaron varios avances en la manipulación de datos, incluyendo la creación de nuevas variables, la imputación de valores faltantes y la eliminación de variables irrelevantes. Estos avances tienen como objetivo mejorar la calidad de las predicciones y, por lo tanto, contribuir a la identificación temprana de solicitantes de crédito de alto riesgo.

Enlace video: https://youtu.be/zx_5X7e7DVo

Enlace competencia: Home Credit Default Risk | Kaggle