INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

PROYECTO IA

JOSE CARLOS ORTIZ PADILLA
1003059949
Ingeniería de Sistemas

ALEXANDER VALENCIA DELGADO
71372112
Ingeniería de sistemas

UNIVERSIDAD DE ANTIOQUIA
MEDELLIN
2023

Home Credit Default Risk

Enlace de la competición en kaggle
https://www.kaggle.com/c/home-credit-default-risk/overview

Enlace del video
https://drive.google.com/file/d/10u3UQzTKPzyCabPePsWB7omZ_VEFru_B/view?usp=sharing

El proyecto Home Credit Default Risk tiene como objetivo predecir la probabilidad de que los solicitantes de crédito incumplan con el pago de sus préstamos. Para lograr este objetivo, es fundamental realizar una manipulación adecuada de los datos antes de su modelado y análisis. En este informe, se describe un avance en la manipulación de datos que puede ayudar a mejorar la calidad de las predicciones.

El primer paso en la manipulación de datos es cargar los archivos de datos de entrenamiento y prueba. Esto se hace utilizando la biblioteca Pandas de Python y la función "read_csv". Luego, se combina los conjuntos de datos en un solo DataFrame utilizando la función "concat". Esta combinación permite una manipulación más fácil y consistente de los datos.

Una vez combinados los datos, se procede a realizar algunas manipulaciones básicas. En este avance, se enfoca en la creación de nuevas variables a partir de variables existentes. En particular, se crean dos nuevas variables: "income_per_person" y "credit_term". La primera variable se calcula dividiendo el ingreso total por el número de personas en la familia, mientras que la segunda variable se calcula dividiendo el monto del préstamo por el plazo en días.

La creación de estas nuevas variables puede mejorar la calidad de las predicciones de incumplimiento de pago. En particular, la variable "income_per_person" puede ser un buen predictor del riesgo de incumplimiento, ya que los hogares con ingresos más bajos pueden tener más dificultades para cumplir con sus pagos. Por otro lado, la variable "credit_term" puede ser útil para identificar a los solicitantes que solicitan plazos de pago demasiado cortos o largos, lo que puede indicar un mayor riesgo de incumplimiento.

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.

Otro avance importante en la manipulación de datos es la imputación de valores faltantes. Los datos faltantes son comunes en cualquier conjunto de datos, y pueden ser problemáticos para los

modelos de predicción si no se manejan adecuadamente. En este proyecto, se utilizó una técnica de imputación por media y mediana para completar los valores faltantes en el conjunto de datos.

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.

En conclusión, el proceso de manipulación de datos es esencial en la creación de modelos de predicción de riesgo crediticio, como el proyecto Home Credit Default Risk. 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.

Es importante tener en cuenta que la manipulación de datos debe ser cuidadosa y basada en el conocimiento experto del dominio. Se deben aplicar técnicas apropiadas y evaluar continuamente la calidad y eficacia de las técnicas utilizadas. Además, la validación adecuada del modelo debe realizarse para garantizar que las predicciones sean precisas y confiables.

En resumen, los avances en la manipulación de datos presentados en este informe son cruciales para 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, lo que a su vez puede ayudar a reducir el riesgo de impago de créditos y mejorar la rentabilidad de la compañía.