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 entrega final
https://www.youtube.com/watch?v=4X0uY4h2u0I

Este informe tiene como objetivo proporcionar una visión general y un análisis detallado de la competición "Home Credit Default Risk" de Kaggle. En esta competición, los participantes se enfrentan al desafío de desarrollar modelos de aprendizaje automático para predecir la probabilidad de incumplimiento de pago por parte de los clientes de Home Credit.

El contexto de la competición se basa en Home Credit, una institución financiera no bancaria que ofrece servicios de crédito a personas con historial crediticio limitado o nulo. El objetivo de Home Credit es brindar acceso a préstamos y servicios financieros a un segmento de la población que tradicionalmente ha sido excluido de los servicios bancarios convencionales. Sin embargo, evaluar el riesgo crediticio en este grupo de clientes puede resultar desafiante debido a la falta de datos crediticios tradicionales.

En este informe, exploraremos en detalle el conjunto de datos proporcionado para la competición, que incluye una amplia gama de características e información sobre los solicitantes de crédito. Estos datos abarcan aspectos demográficos, antecedentes laborales, historiales de pago, transacciones financieras previas y más. Con base en estos datos, los participantes deben desarrollar modelos predictivos que puedan estimar con precisión la probabilidad de que un cliente incumpla con sus pagos de crédito.

Durante el análisis exploratorio de datos, examinaremos la distribución de la variable objetivo, identificaremos posibles desequilibrios de clases y evaluaremos la calidad y la integridad del conjunto de datos. También investigaremos las relaciones y patrones dentro de los datos, tanto entre las variables como en relación con la variable objetivo, para extraer conocimientos relevantes.

La competición "Home Credit Default Risk" de Kaggle es un desafío en el que deben desarrollar modelos de aprendizaje automático para predecir la probabilidad de que un cliente incumpla con sus pagos de crédito. A continuación, una exploración descriptiva básica del dataset:

Importancia de los archivos:

- application_train.csv: Este archivo contiene los datos de entrenamiento principales, que incluyen información sobre los solicitantes de crédito.

- application_test.csv: Este archivo contiene los datos de prueba, y se utiliza para evaluar el rendimiento de los modelos predictivos.

- bureau.csv: Proporciona datos adicionales sobre los préstamos anteriores de los solicitantes de crédito, si están disponibles.

- bureau_balance.csv: Este archivo contiene información mensual sobre el saldo de deuda de los préstamos anteriores del archivo "bureau.csv".

- previous_application.csv: Proporciona datos sobre las aplicaciones previas de los solicitantes de crédito en Home Credit.

- POS_CASH_balance.csv: Contiene información mensual sobre los saldos de los préstamos anteriores relacionados con créditos en tiendas.

- credit_card_balance.csv: Proporciona datos mensuales sobre los saldos de las tarjetas de crédito de los solicitantes de crédito.

- installments_payments.csv: Contiene información sobre los pagos mensuales de los préstamos anteriores.

- HomeCredit_columns_description.csv: Un archivo que describe el significado de las columnas en los archivos de datos principales.

Carga de datos:
Debes cargar los archivos CSV en tu entorno de programación y familiarizarte con los datos que contienen.

Exploración básica de datos:

- Comienza por visualizar algunas filas de los datos para entender la estructura y los tipos de variables.
- Examina la distribución de la variable objetivo (TARGET) para ver si hay un desequilibrio significativo.

- Calcula y analiza las estadísticas descriptivas de las variables numéricas (media, mediana, desviación estándar, etc.).
- Explora las variables categóricas y observa los diferentes valores únicos y su frecuencia.
- Verifica si hay valores faltantes en el dataset y decide cómo manejarlos.

Análisis más detallado:

- Examina las relaciones entre las variables y la variable objetivo mediante gráficos y medidas de correlación.
- Realiza análisis comparativos entre las características de los clientes que incumplen y los que no.
- Investiga si hay variables altamente correlacionadas y considera la posibilidad de eliminar algunas para evitar la multicolinealidad.

El despliegue de un modelo de aprendizaje automático implica varios desafíos y consideraciones importantes. Aquí hay algunos de los principales retos y aspectos a tener en cuenta:

- Infraestructura y entorno de implementación: Es necesario asegurarse de contar con la infraestructura adecuada para alojar y ejecutar el modelo. Esto puede incluir servidores, recursos de almacenamiento, capacidad de procesamiento y configuraciones de red. Además, se deben considerar los requisitos de software y las dependencias del modelo.

- Escalabilidad: Si se espera un alto volumen de solicitudes y tráfico, es importante diseñar el sistema de despliegue para que sea escalable. Esto implica considerar la capacidad de respuesta del modelo ante múltiples solicitudes simultáneas y cómo escalar los recursos según sea necesario.

- Latencia y tiempo de respuesta: Dependiendo del caso de uso, la latencia y el tiempo de respuesta pueden ser factores críticos. Asegurarse de que el modelo se ejecute de manera eficiente y rápida es importante para garantizar una experiencia de usuario fluida.

- Seguridad: La seguridad de los datos y el modelo es esencial. Se deben tomar medidas para proteger los datos sensibles y asegurar que el acceso al modelo esté restringido y controlado. Además, es importante considerar la seguridad de las comunicaciones y las vulnerabilidades potenciales del sistema.

- Monitoreo y mantenimiento: Una vez que el modelo esté en producción, es crucial establecer un sistema de monitoreo para supervisar su rendimiento y detectar posibles

problemas. Además, se deben tener procedimientos establecidos para realizar actualizaciones, mejoras y mantenimiento periódico del modelo.

- Versionado y control de cambios: Mantener un registro claro de las versiones del modelo y los cambios realizados es fundamental. Esto facilitará la reproducción de resultados, la colaboración y la identificación de problemas en caso de ser necesario revertir a versiones anteriores.

- Cumplimiento normativo y ético: Asegurarse de que el despliegue del modelo cumpla con las regulaciones y políticas aplicables es esencial. También es importante tener en cuenta consideraciones éticas, como la equidad y la transparencia en el uso del modelo.

Estos son solo algunos de los retos y consideraciones clave a tener en cuenta al desplegar un modelo de aprendizaje automático. Cada caso puede tener sus propias particularidades y requerimientos adicionales. Es recomendable realizar pruebas exhaustivas y trabajar en colaboración con expertos en el dominio y en infraestructura para garantizar un despliegue exitoso y seguro del modelo.
Para la competición seleccionada de Kaggle "Home Credit Default Risk" hay varios retos y consideraciones específicas que debes tener en cuenta durante el despliegue:

- Manejo del desequilibrio de clases: Es probable que el conjunto de datos presente un desequilibrio entre la clase de clientes que incumplen con los pagos y los que no. Esto puede afectar el rendimiento del modelo y la interpretación de las métricas de evaluación. Debes considerar técnicas como el muestreo estratificado, la ponderación de clases o el uso de algoritmos específicos para datos desbalanceados.

- Adaptabilidad del modelo: Una vez que hayas desarrollado tu modelo, es importante asegurarte de que pueda adaptarse fácilmente a nuevos datos o actualizaciones. Esto implica diseñar una estructura modular y flexible, que permita la incorporación de nuevos datos y la reentrenamiento del modelo de manera eficiente.

- Eficiencia computacional: Los modelos de aprendizaje automático pueden ser computacionalmente intensivos, especialmente si se utilizan algoritmos complejos o si el conjunto de datos es grande. Asegúrate de que tu modelo sea eficiente en términos de tiempo de entrenamiento, carga en memoria y predicciones en tiempo real.

- Interpretabilidad: La explicabilidad del modelo es crucial, especialmente en un contexto de crédito y riesgo. Debes asegurarte de que puedas proporcionar explicaciones claras y comprensibles sobre cómo se toman las decisiones de predicción. Considera el uso de

modelos interpretables o técnicas de explicabilidad, como la importancia de características o el análisis de saliencia.

- Actualización y mantenimiento: El mundo del crédito y el riesgo está en constante cambio, por lo que es importante mantener tu modelo actualizado. Considera cómo podrías automatizar la actualización del modelo en función de nuevos datos y cómo gestionar los cambios regulatorios o en las políticas de crédito.

- Cumplimiento normativo y ético: En el contexto de la competición, es crucial asegurarse de que tu modelo cumpla con todas las regulaciones y políticas aplicables, como las leyes de protección de datos. Además, debes considerar las implicaciones éticas y la equidad en la toma de decisiones crediticias.

PRE-PROCESAMIENTO DE DATOS:

- Cargamos los datos en un entorno de programación utilizando una biblioteca como Pandas.

- Exploramos los datos cargados para comprender las columnas y los tipos de datos.

- Realizamos limpieza de datos, como eliminar duplicados, tratar valores faltantes y corregir errores.

- Codificamos variables categóricas para convertirlas en representaciones numéricas adecuadas para el modelado.

- Normalizamos o escalar las variables para asegurarse de que tengan una escala comparable.
- Realizamos selección o ingeniería de características para identificar las más relevantes o crear nuevas características.

MODELOS SUPERVISADOS:

- Separamos las características (variables independientes) de la variable objetivo (variable dependiente).

- Dividimos los datos en conjuntos de entrenamiento y prueba.

- Seleccionamos un modelo supervisado adecuado para el problema, como regresión logística, árboles de decisión o redes neuronales.

- Entrenamos el modelo utilizando los datos de entrenamiento.

- Realizamos predicciones utilizando el modelo entrenado y los datos de prueba.

- Evaluamos el rendimiento del modelo utilizando métricas apropiadas, como precisión, recall o error cuadrático medio.

MODELOS NO SUPERVISADOS:

- Seleccionamos características relevantes para el modelo no supervisado.

- Filtramos los datos según las características seleccionadas.

- Elegimos un modelo no supervisado adecuado, como K-means para el agrupamiento.

- Ajustamos el modelo utilizando los datos filtrados.

- Obtuvimos las etiquetas o grupos resultantes del modelo.

- Analizamos y visualizamos los grupos obtenidos para obtener información sobre las estructuras o patrones ocultos en los datos.

RESULTADOS, MÉTRICAS Y CURVAS DE APRENDIZAJE:

- Evaluamos los resultados de los modelos utilizando métricas apropiadas, como precisión, recall, error cuadrático medio o coeficiente de determinación.

- Comparamos los resultados de diferentes modelos o configuraciones para determinar el mejor rendimiento.

- Utilizamos curvas de aprendizaje para visualizar el rendimiento del modelo en función del tamaño del conjunto de datos de entrenamiento y la complejidad del modelo.

- Iteramos y ajustamos los modelos, según sea necesario, para mejorar los resultados y la eficacia del modelo.

RETOS Y CONSIDERACIONES DE DESPLIEGUE

1. Datos desbalanceados: El conjunto de datos puede presentar un desequilibrio entre las clases de la variable objetivo. En este caso, la cantidad de clientes que incumplen con los pagos puede ser significativamente menor que aquellos que cumplen. Esto puede afectar la capacidad del modelo para aprender patrones y realizar predicciones precisas.

2. Pre procesamiento complejo: El conjunto de datos puede contener una amplia variedad de características, algunas de las cuales pueden estar incompletas o requerir una limpieza y transformación exhaustiva. Fue necesario realizar un pre procesamiento adecuado, como tratar los valores faltantes, manejar características categóricas y normalizar las variables numéricas.

3. Selección de características relevantes: Dado que el conjunto de datos puede contener una gran cantidad de características, fue crucial identificar aquellas que tuvieran mayor relevancia para predecir el incumplimiento crediticio. Realizamos un análisis detallado y seleccionamos cuidadosamente las características más informativas para mejorar la precisión del modelo y reducir la complejidad.

4. Elección del modelo adecuado: Consideramos varios modelos supervisados, como regresión logística, árboles de decisión y redes neuronales, evaluando su capacidad para manejar el desequilibrio de clases y proporcionar una buena precisión en la predicción del incumplimiento crediticio. La elección del modelo adecuado fue crucial para obtener resultados satisfactorios.

5. Validación cruzada y ajuste de hiperparámetros: Para evitar el sobreajuste y evaluar correctamente el rendimiento de nuestros modelos, utilizamos técnicas de validación cruzada y ajustamos los hiperparámetros de los modelos. Esto implicó encontrar un equilibrio entre la complejidad del modelo y su capacidad para generalizar correctamente a nuevos datos.

6. Interpretación de los resultados: Una vez que obtuvimos los resultados del modelo, fue importante interpretar las predicciones y comprender los factores que influyen en el incumplimiento crediticio. Esto nos permitió extraer información valiosa para tomar decisiones informadas y ofrecer una explicación clara sobre los factores determinantes a los interesados.

CONCLUSIONES

Como equipo, hemos llegado a varias conclusiones importantes a partir de este proyecto de Home Credit Default Risk:

Importancia del preprocesamiento de datos: El preprocesamiento de datos desafiantes y complejos fue fundamental para obtener resultados precisos. Pasamos una cantidad significativa de tiempo limpiando y transformando los datos, lo cual tuvo un impacto directo en la calidad de los modelos finales.

Selección de características significativas: Identificamos las características más relevantes para predecir el incumplimiento crediticio. Al realizar una cuidadosa selección de características, pudimos mejorar la precisión de los modelos y reducir la complejidad.

Desafíos del desequilibrio de clases: La presencia de un desequilibrio entre las clases de la variable objetivo presentó un desafío adicional. Tuvimos que implementar técnicas de manejo de clases desequilibradas, como muestreo estratificado o ajuste de pesos, para obtener predicciones más precisas y evitar sesgos.

Elección del modelo adecuado: Evaluamos y probamos varios modelos supervisados para determinar cuál era el más adecuado para nuestro problema. Encontramos que algunos modelos tenían un mejor desempeño que otros en términos de precisión y capacidad para manejar los desafíos específicos de los datos.

Interpretación de los resultados: Además de obtener predicciones precisas, fue importante comprender y explicar los factores clave que influyen en el incumplimiento crediticio. Esto nos permitió brindar información valiosa a los interesados y tomar decisiones más informadas en futuras estrategias de gestión de riesgos crediticios.

Aprendizaje continuo: A lo largo del proyecto, nos dimos cuenta de la importancia del aprendizaje continuo. Exploramos nuevas técnicas, experimentamos con diferentes enfoques y nos mantenemos actualizados con los avances en el campo del aprendizaje automático. Esta experiencia nos ha ayudado a crecer profesionalmente y nos ha motivado a seguir mejorando en proyectos futuros.

En resumen, este proyecto nos ha permitido adquirir habilidades en el pre procesamiento de datos complejos, la selección de características, la elección de modelos y la interpretación de resultados. Además, hemos comprendido la importancia de abordar desafíos específicos, como el desequilibrio de clases, y hemos experimentado el proceso iterativo y colaborativo que implica trabajar en un proyecto de aprendizaje automático.