Plan Overview

A Data Management Plan created using DMPTool

Title: Predicción de perdida de clientes en tarjetas de crédito

Creator: Cristhian Barbosa montero

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: Liliana Catalina Briceño Casadiegos

Data Manager: Diego Alejandro Peñaloza Mayorga

Project Administrator: Santiago Pulido Vela

Contributor: Cristhian David Barbosa Montero

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

El proyecto busca desarrollar un modelo predictivo capaz de identificar a los clientes de un banco con mayor probabilidad de abandonar sus servicios de tarjeta de crédito. Para ello, se utilizará un conjunto de datos que incluye la información demográfica y financiera de aproximadamente 10,000 clientes. A través del uso de técnicas de Machine Learning, se busca identificar patrones en los datos de los clientes que permitan predecir la probabilidad de que dejen de utilizar el producto (probabilidad de Churn). El resultado otorgará al banco un sistema que le permita implementar acciones efectivas destinadas a retener los clientes en riesgo de abandono, lo que contribuiría a aumentar la tasa de retención y mejorar la satisfacción del cliente.

Start date: 10-20-2024

End date: 12-01-2024

Last modified: 10-20-2024

Predicción de perdida de clientes en tarjetas de crédito

Data Collection

What data will you collect or create?

Se utilizarán datos existentes proporcionados por el banco, que incluyen 18 características por cliente, como edad, ingresos, estado civil, límites de crédito, historial de transacciones y métricas de uso de la tarjeta. No se planea recolectar nuevos datos. Además, se generarán nuevas variables derivadas de las ya existentes, como razones de cambio de comportamiento entre trimestres, tasas de utilización de crédito, y agrupaciones basadas en categorías.

How will the data be collected or created?

Los datos ya han sido recolectados por el banco en el curso de sus operaciones y serán proporcionados en forma de archivo. Se realizarán transformaciones y limpieza de datos, como la codificación de variables categóricas y el balanceo de clases mediante técnicas como SMOTE. Los datos también serán normalizados y estandarizados según sea necesario para los modelos predictivos.

Documentation and Metadata

What documentation and metadata will accompany the data?

Se creará documentación que incluirá:

- Descripción de las variables originales y derivadas.
- Metodología de preprocesamiento, incluyendo los pasos de limpieza y transformación.
- Información sobre las técnicas de machine learning utilizadas.
- Diccionario de datos, con definiciones claras de cada variable.

La metadata acompañará los datos, explicando el propósito de cada campo y el tipo de preprocesamiento aplicado.

Ethics and Legal Compliance

How will you manage any ethical issues?

Para abordar las preocupaciones éticas, se deben tener en cuenta las siguientes consideraciones:

- **Anonimización de datos:** asegurar la anonimización de los datos personales sensibles, eliminando identificadores directos como nombres o números de cuenta.
- **Cumplimiento normativo:** se deben respetar todas las normativas de privacidad aplicables.
- **Transparencia:** fomentar la transparencia en el proceso de análisis, mediante la obtención del consentimiento informado de los participantes, asegurando que conocen como se utilizarán sus datos y los resultados obtenidos, promoviendo un uso responsable de la información.

- **Capacitación del equipo:** proporcionar capacitación a todos los miembros del equipo sobre la ética en el manejo de datos y la su importancia en la aplicación del proyecto.
- **Uso responsable de datos:** el desarrollo del proyecto se utilizará como herramienta de apoyo para abordar la problemática y no como la solución definitiva. Por lo tanto, los resultados del modelo, que indican probabilidad de deserción, no deben usarse como la única base para tomar decisiones para la retención de clientes. Es fundamental interpretar los resultados con cautela, considerando el contexto y otros factores relevantes antes de implementar cualquier estrategia de retención.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

El banco es el propietario de los datos originales y posee los derechos de propiedad intelectual sobre ellos. Sin embargo, los modelos predictivos y los resultados del análisis serán propiedad del equipo investigador o de la entidad que lo financie, según los acuerdos establecidos. Se formalizarán acuerdos de uso de los datos, asegurando que los mismos no se utilicen fuera del ámbito de este proyecto sin permiso.

Storage and Backup

How will the data be stored and backed up during the research?

Los datos serán almacenados de forma segura en servidores con acceso restringido. Se realizará un respaldo en los servidores de almacenamiento en la nube y en servidores locales de manera periódica. Se tendrá como herramienta de administración y control DVC (Data versión Control), el cual permite aplicar buenas prácticas de control de versiones para los conjuntos de datos y los scripts. El proveedor en la nube será Amazon.

How will you manage access and security?

El acceso a los datos estará restringido únicamente a los miembros del equipo del proyecto. Los permisos serán controlados, garantizando que solo las personas autorizadas puedan modificar o acceder a los datos.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Los datos transformados, los modelos predictivos entrenados y los resultados del análisis tendrán un valor a largo plazo. Estos incluyen los modelos entrenados, los conjuntos de datos preprocesados, las variables derivadas, la documentación y metadatos, así como los resultados de los análisis que brinden insumos valiosos y puedan ser reutilizados en estudios futuros. Los datos originales proporcionados por el banco no se compartirán externamente debido a las restricciones de privacidad. Sin embargo, los resultados agregados y de análisis pueden ser compartidos con el banco y demás partes interesadas para el desarrollo de estrategias.

Además, se establecerán políticas claras de retención de datos para asegurar que se mantengan solo aquellos datos necesarios para el análisis y desarrollo futuro, minimizando el riesgo asociado al

almacenamiento de información sensible.

What is the long-term preservation plan for the dataset?

Los datos derivados y los resultados del análisis se almacenarán en repositorios de datos seguros y en cumplimiento de las normativas de protección de datos, de acuerdo con las políticas de preservación de datos de la institución investigadora. Los modelos finales se mantendrán disponibles para su reutilización o auditoría mediante almacenamiento en sistemas de control de versiones como Git.

Data Sharing

How will you share the data?

Los resultados del análisis, los scripts de preprocesamiento y los modelos entrenados se compartirán a través de repositorios abiertos como GitHub, junto con la documentación completa de los detalles de proyecto.

Los datos originales por el banco no se compartirán públicamente debido a restricciones de privacidad. Sin embargo, los datos transformados y anonimizados podrían estar disponibles bajo condiciones controladas.

Finalmente, la comunicación de los resultados se realizará mediante una presentación del equipo, donde se discutirán las conclusiones y hallazgos del proyecto al finalizar el curso.

Are any restrictions on data sharing required?

Sí, los datos originales proporcionados por el banco no se compartirán públicamente debido a las normativas de privacidad y confidencialidad. Solo se compartirán los datos transformados, y siempre de manera anonimizada. Cualquier acceso a los datos deberá ser controlado y autorizado por el banco.

Responsibilities and Resources

Who will be responsible for data management?

Todo el equipo será responsable del manejo de datos, como data managers se encargarán de todo el tratamiento, preprocesamiento, uso y respaldo de la información proporcionada. Teniendo en cuenta que el manejo de datos es un esfuerzo colaborativo, todos los miembros del equipo participarán activamente del proceso, así como el aseguramiento del cumplimiento de los protocolos de seguridad y acceso.

What resources will you require to deliver your plan?

Para implementar el plan de manera efectiva, se necesitarán los siguientes recursos:

- **Infraestructura de almacenamiento:** se requieren servidores locales y soluciones de almacenamiento en la nube para almacenar los datos, gestionarlos y garantizar copias de seguridad.

- **Herramientas de machine learning:** se debe contar con software como Python, junto con bibliotecas como pandas y scikit-learn, además de herramientas de control de versiones como Git y visualización de datos.
- **Seguridad de datos:** implementar sistemas de encriptación y control de acceso para proteger la privacidad e integridad de los datos.
- **Recursos humanos:** formar un equipo que incluya un científico de datos, un ingeniero de software y un gerente de producto, así como un equipo legal para abordar cuestiones como el manejo de los derechos de propiedad intelectual y los acuerdos de uso de datos.
