Universidad de la Ciudad de Aguascalientes

Caso 5

Contexto: Una empresa de telecomunicaciones desea desarrollar un modelo de predicción de (abandono de clientes), utilizando datos de navegación, historial de llamadas y datos demográficos de sus usuarios.

Perspectiva de la Ley de Privacidad

- Recopilación de datos: ¿La empresa cuenta con el consentimiento explícito de los usuarios para recopilar y utilizar sus datos con fines de análisis?
- Almacenamiento de datos: ¿Los datos se almacenan de forma segura y encriptada para evitar accesos no autorizados?
- Uso de datos: ¿El uso de los datos se limita a los fines para los que se recolectaron (data set para predicción)?
- Compartir datos: ¿La empresa comparte los datos con terceros (por ejemplo, proveedores de servicios en la nube)? Si es así, ¿se han implementado medidas de seguridad adecuadas para proteger la información?
- Derechos de los usuarios: ¿Los usuarios tienen derecho a acceder, corregir o eliminar sus datos personales? ¿La empresa cuenta con mecanismos para ejercer estos derechos?

Gia de la Ley de Privacidad en el proyecto

- Diseño del proyecto: Desde el inicio, el equipo de ciencia de datos debe considerar los principios de la Ley de Privacidad al diseñar el proyecto. Esto implica:
- Minimización de datos: Recolectar solo los datos estrictamente necesarios para el modelo.
- Anonimización y seudonimización: Transformar los datos para dificultar la identificación de los individuos.
- Consentimiento informado: Obtener el consentimiento explícito de los usuarios antes de recopilar y utilizar sus datos.
- Aprendizaje federado: Entrenar modelos de machine learning en los dispositivos de los usuarios, sin compartir los datos en bruto.
- Diferencial privacy: Agregar ruido a los datos para proteger la privacidad de los individuos.
- Homomorphic encryption: Realizar cálculos en datos cifrados, sin descifrarlos.
- Implementación de medidas de seguridad: Es fundamental implementar medidas de seguridad robustas para proteger los datos:

Estrategia de Encriptación

- Control de acceso: Limitar el acceso a los datos a personal autorizado.
- Auditorías: Realizar auditorías periódicas para garantizar el cumplimiento de la normativa.
- Transparencia: La empresa debe ser transparente sobre cómo utiliza los datos de sus clientes. Esto implica
- Aviso de privacidad: Proporcionar un aviso de privacidad claro y conciso que explique cómo se recopilan, utilizan y protegen los datos.
- Informes de impacto: Evaluar el impacto de los modelos de machine learning en la privacidad de los individuos.

Aportaciones de la Ley de Privacidad:

- Confianza de los clientes: Los clientes estarán más dispuestos a compartir sus datos si saben que están protegidos.
- Evitar sanciones: El incumplimiento de la normativa puede acarrear sanciones económicas y reputacionales.
- Innovación responsable: La Ley de Privacidad fomenta la innovación en el campo de la ciencia de datos, al tiempo que garantiza la protección de los derechos de los individuos.

Preguntas:

¿Qué otras técnicas de privacidad podrían aplicarse en este caso?

1. Tokenización: Transformar información personal identificable en un conjunto de caracteres únicos (tokens) que no tienen significado por sí mismos. Los tokens se pueden intercambiar o almacenar de manera segura, mientras que los datos originales se mantienen fuera de los sistemas principales.
2. Cifrado extremo a extremo: Implementar cifrado durante todo el ciclo de vida de los datos, desde la recolección hasta el almacenamiento y el procesamiento, garantiza que los datos estén protegidos en cada etapa, especialmente en entornos de almacenamiento en la nube.
3. División de datos (Split Learning): Esta técnica distribuye los datos entre varios servidores, donde cada uno entrena una parte del modelo sin tener acceso a los datos completos. De esta forma, se reduce el riesgo de que un solo punto de acceso comprometa la privacidad de los usuarios.
4. Cómputo multipartito seguro (MPC): Permite que varias partes realicen cálculos conjuntos en los datos sin tener que compartirlos entre sí. Esta técnica garantiza que los datos permanezcan privados, incluso cuando diferentes equipos o empresas colaboran en el desarrollo de modelos.
5. Gestión del ciclo de vida de los datos: La empresa debe implementar un enfoque

integral para gestionar la vida útil de los datos, asegurando que los datos no se retengan más allá del tiempo necesario y que sean eliminados de forma segura una vez cumplido su propósito.

¿Cómo se pueden equilibrar los intereses comerciales con la protección de la privacidad?

1. Consentimiento Informado y Transparente: Para equilibrar los intereses comerciales, es clave obtener el consentimiento explícito de los usuarios de manera transparente, explicando los beneficios que el análisis de datos traerá tanto a la empresa como a los clientes (mejores servicios, recomendaciones personalizadas, etc.).
2. Minimización de Datos: Recopilar solo los datos estrictamente necesarios para los fines del proyecto de predicción de abandono de clientes. Esto no solo protege la privacidad de los usuarios, sino que también optimiza los recursos y reduce riesgos para la empresa.
3. Anonimización y Seudonimización: Aplicar estas técnicas en los datos personales utilizados para el análisis permite que la empresa utilice la información sin comprometer la identidad de los usuarios, lo que ayuda a proteger la privacidad al tiempo que se obtienen insights comerciales valiosos.
4. Diseño Proactivo de Privacidad: Incluir la privacidad desde el diseño ("privacy by design") en todos los aspectos del proyecto, asegurando que las prácticas de privacidad no sean una barrera para la innovación, sino una forma de gestionar responsablemente los datos. Esto no solo protege a los usuarios, sino que también evita sanciones y construye confianza.
5. Transparencia y Confianza del Usuario: Al ser transparente sobre cómo se usan y protegen los datos, la empresa genera confianza con sus clientes, lo que aumenta la probabilidad de que estos acepten compartir más datos. Esto no solo beneficia al proyecto de predicción, sino que también fortalece la relación comercial.

¿Qué desafíos éticos plantea el uso de datos personales en la ciencia de datos?

1. La empresa debe ser cuidadosa al recolectar datos de navegación, historial de llamadas y datos demográficos. A pesar de obtener el consentimiento, los usuarios podrían no estar plenamente conscientes de la magnitud de los datos recolectados, lo que plantea cuestiones éticas sobre la transparencia y el uso de esos datos para fines más allá de lo explícitamente comunicado.

2. Aunque se utilicen técnicas como la anonimización, siempre existe el riesgo de que los usuarios puedan ser reidentificados a partir de patrones complejos o combinaciones de datos. Esto plantea un desafío ético sobre si los datos "anonimizados" son realmente seguros.
3. El consentimiento informado puede ser complejo de lograr, especialmente si los usuarios no entienden completamente las implicaciones de cómo se utilizarán sus datos en modelos de machine learning. Asegurar que el consentimiento sea informado y comprensible es un reto ético.
4. El uso de datos demográficos puede introducir sesgos en los modelos predictivos. Si no se controlan estos sesgos, los algoritmos pueden discriminar a ciertos grupos, lo que plantea serios problemas éticos y de equidad en el análisis de datos.
5. Los datos recolectados podrían ser utilizados para otros fines comerciales o de monitoreo sin el consentimiento explícito del usuario, lo que plantea un dilema ético sobre el uso justo y transparente de la información personal.

Privacidad Diferencial: Además de añadir ruido a los datos para proteger la privacidad, esta técnica permite realizar análisis estadísticos sobre los datos sin comprometer la identidad de los usuarios. Esto sería útil para entrenar el modelo de predicción de abandono sin exponer información personal sensible.

La privacidad diferencial es un estándar para los cálculos de datos que limitan la información personal que revela un resultado. La privacidad diferencial suele utilizarse para compartir datos y permitir inferencias sobre grupos de personas, a la vez que evita que alguien conozca información sobre un individuo.

https://cloud.google.com/bigquery/docs/differential-privacy?hl=es-419

Aprendizaje Federado: En lugar de recolectar todos los datos en un solo lugar, esta técnica permite entrenar modelos directamente en los dispositivos de los usuarios, de modo que los datos brutos nunca salen del dispositivo. Solo se envían actualizaciones del modelo, lo que protege la privacidad de los datos individuales.

https://la.blogs.nvidia.com/blog/que-es-el-aprendizaje-federado/

Encriptación Homomórfica: Esta técnica permite realizar cálculos y entrenar modelos sobre datos encriptados sin necesidad de descifrarlos. Así, se asegura que los datos permanezcan cifrados durante todo el proceso, lo que garantiza su seguridad incluso durante el análisis.

https://ciberseguridad.com/guias/prevencion-proteccion/criptografia/cifrado-homomorfico/

Tokenización: Transformar información personal identificable en un conjunto de caracteres únicos (tokens) que no tienen significado por sí mismos. Los tokens se pueden intercambiar o almacenar de manera segura, mientras que los datos originales se mantienen fuera de los sistemas principales.