Anomaly detection in insurance financial information

Seminario de Analítica y Ciencia de Datos

# Resumen Descriptivo del Proyecto

Desarrollar un modelo de aprendizaje automático no supervisado, que permita identificar las etiquetas para reconocer transacciones e información financiera inusual en el libro diario (ver entregable 1 con correcciones)

## Problema.

* 1. **¿Cuál es mi problema alrededor de mi proyecto?**

La compañía que brinda los datos para este proyecto es una empresa en búsqueda de cotizar en la bolsa de nueva York, y para tal fin debe garantizar a los accionistas de la compañía que la información financiera es Veraz y confiables. Para esto debe identificar en el libro diario donde se registran las transacciones a nivel contable cuales datos son inusuales y deben investigarse a profundidad para identificar la causa y validar que esta no sea por fraude o errores en los estados financieros.

* 1. **¿Por qué se manifiesta el problema?**

El recurso en el área contable de la organización es limitado. Por lo tanto, no hay suficiente recursos humanos ni herramientas que permitan hacer eficiente el análisis y a la identificación de transacciones inusuales que se registran en el libro diarios de la compañía.

* 1. **¿De quién es el problema?**

El problema es del área de gestión contable integral, quienes por el alto volumen de operaciones son pueden identificar las transacciones inusuales. También representa un problema para el área de auditoria y control interno, ya que no pueden identificar si hubo transaccione inusuales que representan una alerta para los estados financieros de la compañía.

## Datos.

* 1. **Describa la fuente de datos**

La fuente de datos es SAP, ya que es la aplicación que utiliza la compañía para registrar las operaciones del libro diario.

* 1. ¿**Cuál es la calidad que posee dicha fuente?**
* Es una fuente de buena calidad ya que el sistema tiene restricciones como: Que todos los campos estén completos.
* Que las cuentas contables y su descripción sean las registradas por el área contables
* Que los valores una vez ingresados no puedan ser modificados, además la mayor parte de la información llega por interfaz automática proveniente de las tablas de las bases de datos de los aplicativos de negocio en Oracle.
  1. **¿Es accesible y disponible?**

La información es accesible y disponible para las personas de la compañía autorizadas para acceder a la misma, para efectos del trabajo de grado estamos autorizados para el uso de la información.

## Hipótesis.

* 1. **¿Qué es lo que pretende probar?**

A través de la aplicación modelos de machine learning no supervisados como el autoencoder LSTM se pueden etiquetar las transacciones, usuales e inusuales del libro Diario.

* 1. **¿Cuáles pueden ser las respuestas esperadas a lo que pretende probar?**

La respuesta esperada es que después de la preparación del modelo, una vez se arrojen los datos inusuales se puedan identificar y definir los umbrales de anomalías y se tengan las etiquetas necesarias para entrenar un modelo supervisado.

* 1. **¿Qué acciones debo tomar para cada respuesta anterior?**

1. Realizar análisis exploratorio
2. Dividir los datos en conjuntos y subconjuntos, con el fin de perder hacer validación cruzada.
3. Determinar el umbral de reconstrucción.
4. Etiquetar como anomalías los datos comuna error de reconstrucción alto.
5. Realizar un análisis de las anomalías, este análisis con el conocimiento contable y financiero e incluso de sensibilidad de cifras financieras para saber si realmente es una anomalía o no. **Comprobar hipótesis**

## Solución.

* 1. **Tipo. (¿Qué tipo de algoritmo de ML es? ¿Qué técnica es?**

Los autoencoders son una técnica de aprendizaje no supervisado, se utilizan para poder aprender representaciones codificadas eficientes de los datos de entrada, realizando reducción de dimensionalidad de los datos. En cuanto a los LSTM (long-short-term- memory), son un tipo especial de red neuronal recurrente (RNN) capaces de aprender y recordar dependencias de largo plazo en los datos de secuencia, muy útil para series temporales. Combinando las dos técnicas es posible aprender representaciones útiles d ellos datos de secuencia y reconstruir las secuencias de entrada originales. Siendo útil para la detección de anomalías en series temporales.

.

* 1. **¿Cómo debe ser implementado?**

1. **Preprocesamiento de datos:** Recopilar y limpiar datos del libro mayor financiero.
2. **División de datos:** Dividir los datos en conjuntos de entrenamiento, validación y prueba.
3. **Diseño del modelo de autoencoder LSTM:** Definir técnicas de auto regularización dependiendo del análisis exploratorio. Utilizando Keras
4. **Entrenamiento del modelo:** Entrenar el modelo utilizando los datos de entrenamiento. (Machine Learning Mastery , 2020)
5. **Validación del modelo:** Realizar análisis adicionales, como la prueba Grubbs, en los datos de validación para evaluar si el modelo está detectando correctamente las anomalías.
6. **Prueba del modelo:** Probar el modelo utilizando los datos de prueba y realizar análisis adicionales para evaluar su desempeño.
7. **Ajuste del umbral de detección de anomalías:** Ajustar el umbral de detección de anomalías para equilibrar la tasa de falsos positivos y falsos negativos.
8. **Despliegue del modelo.**
   1. **¿Qué resultados se esperan?**

* El modelo debería ser capaz de detectar e identificar anomalías conocidas que se han observado durante el entrenamiento.
* El modelo también puede ser capaz de detectar anomalías desconocidas que no se han visto durante el entrenamiento.
* Se espera que el modelo tenga un buen rendimiento en términos de precisión, recall y F1-score en la detección de anomalías, sea capaz de identificar correctamente la mayoría de las anomalías sin generar demasiados falsos positivo.
* Se espera que el modelo pueda detectar anomalías de manera rápida y eficiente.

## KPI’s

* 1. **¿Cómo espera evaluar el modelo?**

Con la matriz de confusión, que ayuda a visualizar a la relación de falsos positivos y falsos negativos del modelo. Conociendo la Exactitud del modelo (Acurracy), la tasa de errores, la sensibilidad del modelo (recall), el F1 score y realizar la curva ROC.

* 1. **¿Qué métricas debe usar?**

1. **Precisión:** Verdaderos positivos y total clasificados positivos.
2. **Recall:**  Verdaderos positivos y total positivos
3. **Especificidad:** Verdaderos negativos y total negativos.
4. **F1- Score:** Precisión y recall
5. **Curva ROC:** Especificidad y recall

## Actores.

* 1. **¿Quién es su cliente?**

El área contable, de control financiero y auditoria de la compañía.

* 1. **¿Quiénes son los interesados (Stakeholders)?**

El área contable y de control financiero.

* 1. **¿Quién usará la solución?**

Coordinador de operaciones contable

* 1. **¿A quién impactará?**

Impacta todas las áreas que generan transacciones y estas se registran en el libro diario, es decir posiblemente impacte todas las áreas de la compañía que generan información financiera.

1

[https://medium.com/@leandroscarvalho/data-product-canvas-a-practical-framework-for-building-high-](https://medium.com/%40leandroscarvalho/data-product-canvas-a-practical-framework-for-building-high-) performance-data-products-7a1717f79f0

# Rúbricas

R01: Completitud y consistencia: el estudiante realiza una entrega del momento evaluativo con todos los elementos desarrollados, manteniendo una relación coherente entre los elementos solicitados. El estudiante usa una consistencia argumentativa que incluye el uso de referencias cuando hay mérito de autoría entre otros.

R02: Pertinencia: el proyecto del estudiante está dentro del ámbito de la analítica y la ciencia de datos y es evidente el desarrollo de cada uno de los elementos del Data Product Canvas

R03: Solución: el estudiante es capaz de responder a las preguntas del Data Product Canvas con la consistencia necesaria para relacionarla con el proyecto que desarrolla. Hay una relación evidente entre ellos. Las herramientas de MAchine Learning son una solución viable dentro de la descripción de la necesidad y objetivos del proyecto.

R04: Datos: los datos sugeridos son pertinentes al proyecto y al alcance delimitado en el momento evaluativo anterior.