PEC 1: DEFINICIÓN Y PLANIFICACIÓN DEL TRABAJO FINAL

Propuesta de título: Desarrollo de un modelo predictivo sobre fraude en transacciones basadas en el uso de tarjetas de crédito.

Santiago Domínguez Collado

1. Introducción

1.1 Contexto y justificación del Trabajo

Punto de partida del trabajo (Cuál es la necesidad a cubrir? Por qué es un tema relevante? Cómo se resuelve el problema de momento?) y aportación realizada (Qué resultado se quiere obtener?)

El trabajo parte de un set de datos que contiene una gran cantidad de transacciones de tarjetas de crédito.

Nuestro objetivo será desarrollar un modelo predictivo que sea capaz de discernir qué transacciones son fraudulentas para así detectar y prevenir actos delictivos como la suplantación de identidad y el robo.

Partimos de unos datos altamente desbalanceados en los que solo el 0.17% de los registros corresponden a transacciones fraudulentas. Mediante un preprocesado exhaustivo balancearemos y prepararemos el set de datos para poder aplicar un modelo que sea capaz de clasificar correctamente las operaciones fraudulentas.

Puesto que es fácil que en estos casos que el modelo siempre muestre una buena precisión o *accuracy,* la calidad de nuestros resultados se medirá principalmente con la métrica de exhaustividad o *recall*, que reflejará qué porcentaje de éxito tenemos en la detección de transacciones fraudulentas respecto al total de las transacciones de este tipo.

En la actualidad este problema es abordado con por las grandes compañías realizando un Análisis de datos exploratorio (EDA), seguido de una reducción de la dimensionalidad y utilizando modelos de Machine Learning para llevar a cabo la clasificación.

Nosotros proponemos replicar este modelo con profesionalidad, justificando todas las decisiones tomadas y estudiando la posibilidad de ampliar la fase de modelaje aplicando modelos de Deep Learning si procede.

1.2 Objetivos del Trabajo

Listado de los objetivos del trabajo

- Realización de la PEC 2: Estado del arte o análisis de mercado del proyecto, cumpliendo con todos sus requisitos.
- Elección justificada del entorno de desarrollo más adecuado para el trabajo, sin extendernos en arquitecturas Big Data que alejen el trabajo del enfoque centrado en la Ciencia de Datos que buscamos.

- Aplicación de un EDA, buscando normalizar los datos que aún no lo estén y analizar las características de las variables. Reduciremos la dimensionalidad si resulta adecuado.
- Creación de un set de datos balanceado en el que se hayan ajustado distribución de clases del conjunto de datos mediante técnicas de sobremuestreo, submuestreo o la combinación de ambas.
- Estudio y aplicación de reducción de la dimensionalidad si esta resulta beneficiosa.
- Desarrollo del mejor modelo de clasificación posible que minimice la métrica *recall.*
- Redacción de la memoria final del trabajo que englobe la realización de todos los objetivos previos y las conclusiones del trabajo.

1.3 Enfoque y método seguido

Indicar cuáles son las posibles estrategias para llevar a cabo el trabajo e indicar cuál es la estrategia elegida (desarrollar un producto nuevo, adaptar un producto existente, …). Valorar porque esta es la estrategia más apropiada para conseguir los objetivos.

La metodología que mejor se adapta a nuestra tarea es la famosa CRISP-DM (del inglés Cross Industry Standard Process for Data Mining).

Esta metodología nos servirá en el desarrollo de nuestro producto, que se dividirá en seis fases:

- Comprensión del negocio: cubierta principalmente por el trabajo de la PEC2 en la que se evalúa el estado del arte del proyecto y se analiza el mercado del proyecto.
- Comprensión de los datos: en la que aterrizamos sobre los datos y entendemos con lo que estamos tratando.
- Preparación de los datos: esta fase se divide en tres etapas.
    - Análisis de datos exploratorios.
    - Normalización y ajuste.
    - Creación de un conjunto de datos balanceado a raíz de conjunto original.
- Modelado: creación de un modelo de clasificación.
- Evaluación: estudio de los resultados del modelo.
- Despliegue: que en nuestro caso se limitará a la redacción de un informe en el que se reflejen los resultados obtenidos.

He escogido esta métrica respecto a otras igual de válidas (como KDD) porque en la definición de esta se describe una naturaleza recurrente entre algunas de sus fases.

En concreto la posible vuelta atrás desde la fase de modelaje a la de preparación de los datos será de vital importancia para corregir errores del modelaje que vengan acarreados de la fase previa.

1.4 Planificación del Trabajo

Descripción de los recursos necesarios para realizar el trabajo, las tareas a realizar y una planificación temporal de cada tarea utilizando un diagrama de Gantt o similar. Esta planificación tendría que marcar cuáles son los hitos parciales de cada una de las PEC.

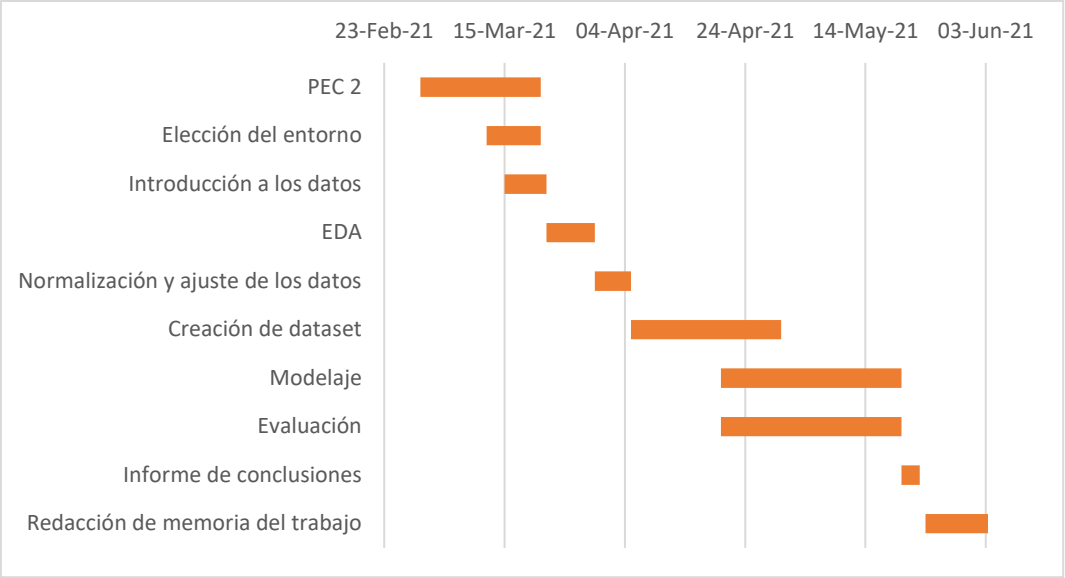
En lo relacionado con los recursos necesarios solo harán falta un ordenador con GPU, el lenguaje Python y librerías que se definirán en el futuro.

A continuación, se explican las tareas a realizar durante este proyecto:

1. PEC 2: en la que se realiza un estudio del estado del arte y de mercado.
2. Elección del entorno: así como de las herramientas a utilizar.
3. Introducción a los datos: recabar información y características del conjunto de datos.
4. Análisis de datos exploratorios.
5. Normalización y ajuste de los datos: realización de ajustes necesarios a los datos.
6. Creación de un conjunto de datos balanceado a raíz de conjunto original.
7. Modelaje.
8. Evaluación.
9. Informe de conclusiones.
10. Redacción de memoria del trabajo.

La siguiente tabla muestra las fechas estimadas para cada tarea, así como un diagrama de Gantt para ilustrarlo de manera más visual.

Task	Start Date	End Date	Duration
PEC 2	01-Mar-21	21-Mar-21	20
Elección del entorno	12-Mar-21	21-Mar-21	9
Introducción a los datos	15-Mar-21	22-Mar-21	7
EDA	22-Mar-21	30-Mar-21	8
Normalización y ajuste de los datos	30-Mar-21	05-Apr-21	6
Creación de dataset	05-Apr-21	30-Apr-21	25
Modelaje	20-Apr-21	20-May-21	30
Evaluación	20-Apr-21	20-May-21	30
Informe de conclusiones	20-May-21	23-Jun-21	3
Redacción de memoria del trabajo	24-Jun-21	06-Jun-21	13



1.5 Breve sumario de productos obtenidos

No hay que entrar en detalle: la descripción detallada se hará en el resto de capítulos.

- Se presentarán una serie de scripts y notebooks en Python como resultado de todas las fases del proyecto.
- Un archivo ".pb" o ". joblib" con el clasificador ya entrenado.
- Un conjunto de datos balanceado de transacciones de tarjetas de crédito en formato ".csv".
- Un script en Python para poder probar el modelo final, solo con código para comprobar su funcionamiento.
- Una memoria detallando el trabajo realizado durante todo el proyecto

2.6 Breve descripción de los otros capítulos de la memoria

Explicación de los contenidos de cada capítulo y su relación con el trabajo en global.

Capítulo 2. Los datos: capítulo de exploración de los datos en los que se mostrarán características relevantes de los mismos.

Capítulo 3. Transformaciones y nuevo set de datos: se explicarán las transformaciones realizadas (normalización, reducción de dimensionalidad…) y aplicarán técnicas de balanceo para crear un nuevo conjunto de datos con el que poder modelar.

Capítulo 4. Modelaje: apartado en el que se explicará el proceso de desarrollo del modelo, así como la justificación de por qué se ha elegido dicho modelo.

Capítulo 5. Producto final y conclusiones: en este apartado se explicará lo conseguido en el trabajo y se expondrán unas conclusiones sobre el producto final y el proyecto en general.

Capítulo 6. Bibliografía: Bibliografía utilizada en el proyecto.