FRAUDOLENT TRANSACTION CLASSIFICATION

RICCARDO LA MARCA, MATR. 1795030

BIG DATA COMPUTING COURSE A.A. 2021/2022

OVERVIEW

1. **Introduction**
   A brief presentation of the addressed problem
2. **Dataset**
   A brief description of the dataset used in the project
3. **Explore and Feature Engineering**
   How the dataset was modified
4. **Machine Learning Models**
   The ML models and Pipelines applied for the task
5. **Results**
   A description of the results obtained from the previous step

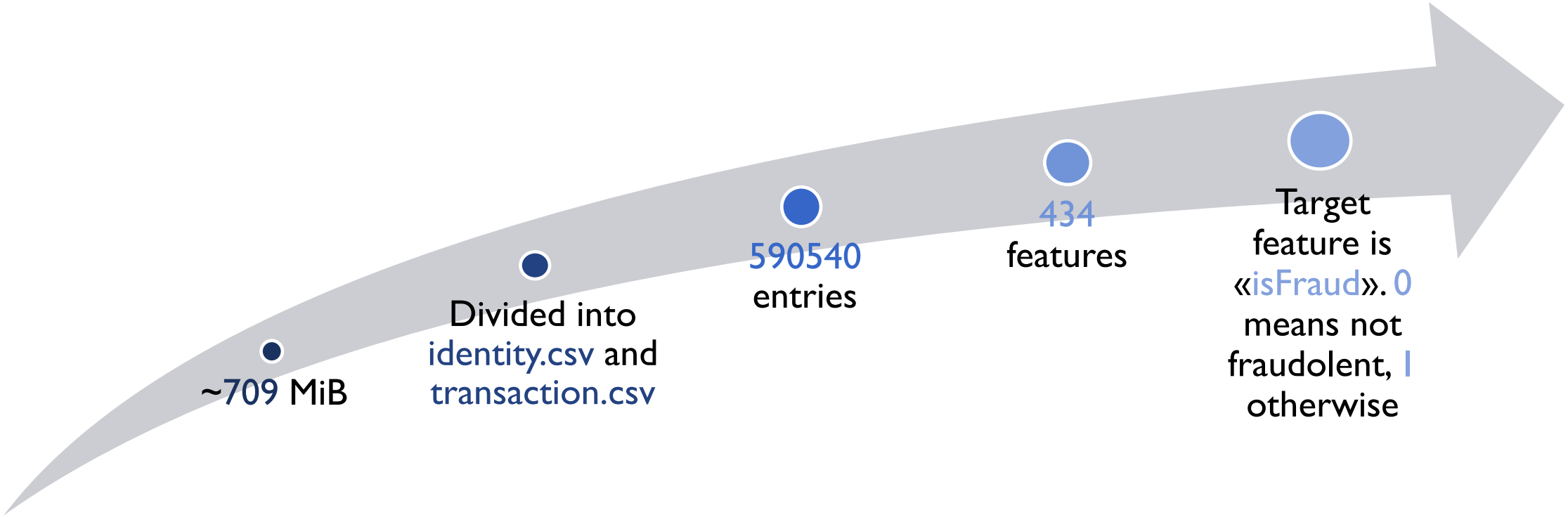
1 ADDRESSED PROBLEM

Financial fraud is a problem that has a huge impact on the financial industry

Credit card fraud detection is a challenge mainly due to 2 problems that it poses

- Both profiles of fraudolent and normal behaviours change
- Usually used datasets are highly skewed

The goal of the task is to create a Machine Learning model that, given a set of samples of fraudolent and not fraudolent transactions, is capable of classifying whether a new transaction is fraudolent or not.



2 THE DATASET

The Dataset is available on Kaggle

3 EXPLORE AND FEATURE ENGINEERING

Merging

- Transaction.csv and identity.csv have been merged together

Label Distributrion

- The dataset is highly unbalanced

Missing Values

- The dataset has a huge number of missing values