


# Project 1 – Sales prediction using ML



The background features a blurred financial chart with a line graph and a bar chart. A blue horizontal line is positioned below the title.

# Introduction

---

- Before we jump further, let's have an end-to-end experience of solving a real-world Machine Learning problem.
- We will try to predict the sales of different stores.
- You have a dataset containing information on stores' sales per day (and some additional info) – your goal is to predict the sales.

# Dataset

---

- Training data (640,841 entries): training set of store sales per day, with bits of information of what happened in that day in that store.
- Real-Life Data (+70k entries): entries without the sales (you need to predict them). This will be used to verify how good your model really is.

# Data Dictionary

Variable	Type	Description	Values/Range
index	Integer	Unique row identifier	0 to N
store_ID	Integer	Unique identifier for each store	1 to 1115
day_of_week	Integer	Day of the week	1 = Monday 2 = Tuesday 3 = Wednesday 4 = Thursday 5 = Friday 6 = Saturday 7 = Sunday
date	Date	Calendar date of the record	Date format
nb_customers_on_day	Integer	Number of customers who visited the store	0 to maximum
open	Binary	Store opening status	0 = Closed 1 = Open
promotion	Binary	Promotional campaign status	0 = No promotion 1 = Promotion active
state_holiday	Categorical	Public holiday indicator	0 = No holiday a = Public holiday b = Easter holiday c = Christmas
school_holiday	Binary	School holiday indicator	0 = Not a school holiday 1 = School holiday
sales	Integer/Float	Daily sales revenue <b>TARGET VARIABLE</b>	0 to maximum

# Deliverables

---

## Expected Delivery:

- “Real-life data set” with an extra column called “sales”, with your predictions. Name this G1.csv (or G2, G3...)
- An expected value of  $R^2$  of performance of your model. Save the number in a file g1\_r2\_prediction.txt (or G2, G3...)
- Your code (Jupyter notebook)
- A document explaining the steps you've followed (EDA, feature engineering, data cleaning, models trained, evaluation, etc.)