Лабораторна робота №9.
Оцінка інформативності ознак за допомогою градієнтного бустінгу

Детальних інструкцій не буде, завдання буде описано тезисно. Рекомендується використати метод Xgboost чи Catboost, та визначити інформативність ознак.

Спочатку налаштуємо доступ до даних на google drive (якщо ви відкриваєте блокнот в google colab, а не на PC) шляхом монтування google drive

In [1]:
```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

Перевіримо шлях до папки з матеріалами лабораторної роботи на google drive. Якщо у вас шлях відрізняється то відредагуйте

In [2]:
```
!ls gdrive/'My Drive'/TEACHING/IntroDataScience/intro_to_data_science/Lab_7_8/
```

```
data
img
lab7_credit_scoring_random_forest.ipynb
lab7_credit_scoring_random_forest.pdf
lab8_quality_metrics_classification.ipynb
lab8_quality_metrics_classification.pdf
lab_9_xgb_flight_delays.ipynb
```

Перемістимо матеріали лабораторної роботи з google drive на віртуальну машину google colab

In [3]:
```
!cp -a gdrive/'My Drive'/TEACHING/IntroDataScience/intro_to_data_science/Lab_7_8/
!ls
```

```
data
gdrive
img
lab7_credit_scoring_random_forest.ipynb
lab7_credit_scoring_random_forest.pdf
lab8_quality_metrics_classification.ipynb
lab8_quality_metrics_classification.pdf
lab_9_xgb_flight_delays.ipynb
sample_data
```

In [4]:
```
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.metrics import roc_auc_score
```

In [5]:
```
train = pd.read_csv('data/flight_delays_train.csv')
test = pd.read_csv('data/flight_delays_test.csv')
```

In [6]:
```
train.head()
```

Out[6]:

	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance	dep_delayed_
0	c-8	c-21	c-7	1934	AA	ATL	DFW	732	
1	c-4	c-20	c-3	1548	US	PIT	MCO	834	
2	c-9	c-2	c-5	1422	XE	RDU	CLE	416	
3	c-11	c-25	c-6	1015	OO	DEN	MEM	872	
4	c-10	c-7	c-6	1828	WN	MDW	OMA	423	

In [7]:
```
test.head()
```

Out[7]:

	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance
0	c-7	c-25	c-3	615	YV	MRY	PHX	598
1	c-4	c-17	c-2	739	WN	LAS	HOU	1235
2	c-12	c-2	c-7	651	MQ	GSP	ORD	577
3	c-3	c-25	c-7	1614	WN	BWI	MHT	377
4	c-6	c-6	c-3	1505	UA	ORD	STL	258

Необхідно за часом вильоту літака, коду авіакомпанії-перевізника, місцю вильоту і прильоту і відстанню між аеропортами вильоту і прильоту зпрогнозувати затримку вильоту більше 15 хвилин.

Завдання 1. Створити вибірку :

- ознаки Distance і DepTime використати без змін;
- створити ознаку "маршрут" на основі Origin та Dest;
- до ознак Month, DayofMonth, DayOfWeek, UniqueCarrier і "маршрут" застосувати OHE-перетворення (LabelBinarizer);

Завдання 2. Побудувати модель і оцінити інформативність ознак :

інформативність ознак?

- побудувати модель на основі xgboost;
- налаштувати гіперпараметри з використанням крос-валідації;
- оцінити інформативність ознак.