Women Data Science Hackathon by Bain & Company

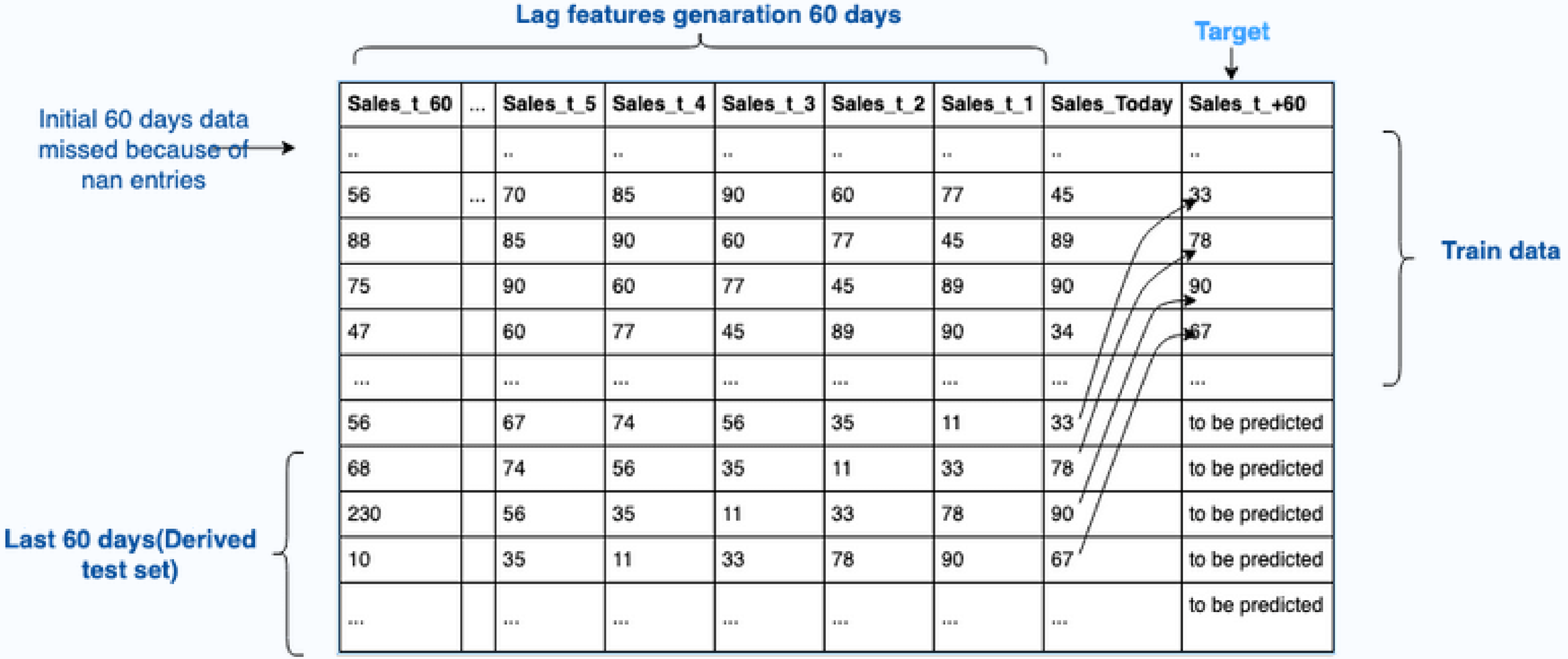
Presented by Lavanya M K

Exploratory Data Analysis

1. The training data contains sales of 600 courses for 882 days. Following is the distribution:
    a. 506 Courses -> 882 days
    b. 93 Courses -> 698 days
    c. 1 Course -> 881 days
2. There are 10 columns excluding target column (Sales) in train data. There are 9 columns in given test data.
3. Course type 'Program' and 'Course' has lesser sales than 'Degree' even though the number of courses for 'Degree' is only 2 and courses for types 'Program' and 'Course' are 288 and 310 respectively.
4. The average sales for Course_Type Degree is more in public holiday where as for other course type it is less in public holiday.
5. Sales pattern for Business Domain is different from other three domains - Software Marketing, Finance & Accounting, Development
6. The sales is always highly influenced by User_Traffic
7. Short_Promotion influenced more for sales across each domain than Long_Promotion

Approach

Since the prediction has to be done for next 60 days for each courses, Previous 60 days of Sales, User_Traffic, Long_Promotion, Short_Promotion lag features are formed including current day's columns to predict for 60th day Sales. Lag feature and derived test feature for a single course is shown in the diagram below



Preprocessing and Feature creation

1. Missing values present in Competition_Metric is imputed with 0 values
2. For each course ID, Lag features are created including previous 60 days of Sales, User_Traffic, Long_Promotion, Short_Promotion.
3. Next 60th day Sales is considered as target column
4. Last 60 rows of each course which corresponds to last 60 days of train data is combined with test data for prediction which acts as lag features for test data
5. Holdout set is created to test the model accuracy

Model Creation

LSTM model is used to predict Sales. With leaky Relu as activation layer. Model architecture is described below

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
cu_dnnlstm_1 (CuDNNLSTM)     (None, 512)               1054720
_________________________________________________________________
leaky_re_lu_1 (LeakyReLU)    (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               262656
_________________________________________________________________
leaky_re_lu_2 (LeakyReLU)    (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               65664
_________________________________________________________________
leaky_re_lu_3 (LeakyReLU)    (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 32)                4128
_________________________________________________________________
leaky_re_lu_4 (LeakyReLU)    (None, 32)                0
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 33
=================================================================
Total params: 1,387,201
Trainable params: 1,387,201
Non-trainable params: 0
_________________________________________________________________
```

Thank you