ML Hackathon

Hackathon Logistics

- What is a Hackathon?
  - 2-day online team competition to solve a machine learning problem

- Agenda for next 2 days
  - Day1– 11:30am :: Kick-off and reveal of problem statement and dataset
  - Day1 – 11:30 am :: Session - How to approach a hackathon?
  - Day1 – 4:00 pm onwards :: Team-wise review session

  - Day2 – 12:00 am onwards :: Team-wise review session
  - Day2 – Midnight :: Deadline to submit presentations

Problem Statement

- Predict if a customer is going to accept a coupon for a particular venue, considering demographic and contextual attributes
  - train data - 10,147 records
  - test data – 2,537 records

You are hired as a data scientist at a leading shopping mall in the country. The shopping mall has tied up with different restaurants/bars to provide discount coupons to all its customers. The coupons increase the footfalls at these restaurants and helps the shopping mall to attract more customers. The organization have been relying simple guidelines to determine what coupons are to be provided to the customers, however the organization feels that they need a more robust model to determine whether a customer will accept the recommended coupon or not to improve the use rate. Organization plans to use a mix of client's details that they have captured to create this model.

You are provided with the historical data of the recommended coupons along with customer details in the previous years and your task is to come up with a model which would be able to predict whether a customer will accept the recommended coupon.

TSW

How to submit the solution for evaluation?

- The dataset provided has three files
  - train.csv (labelled data)
  - test.csv (un-labelled data)
  - sample_submission.csv (submission format)

- Teams need to submit the csv file along with predicted labels in the sample_submission.csv format

- The leaderboard will be published over here

Evaluation

Each team will have to submit their solutions in a PPT format
The teams will be evaluated on following

- **Exploratory data analysis** and insights from the data (25 points)
- **Accuracy** of the model (50 points)
- **Presentation Skills** (25 points)

At the end of the hackathon all teams need to submit

- The final Jupyter Notebook file along with any intermediate processed datasets
- Presentation Files

Questions?

How to approach a Hackathon?

Steps to be followed

- Understand the dataset
  - Total number of features available
  - Target feature and its distribution
  - Data types of different features (number, text, boolean, date, etc)
  - Types of features (numerical and categorical)
  - Missing values
  - Number of unique values in categorical features

Steps to be followed

- Clean / pre-process data
  - Remove or impute missing values
  - Check for outliers and take corrective action
  - Check for any incorrect values
  - Encode categorical / scale numerical features

Steps to be followed

- Explore the data
  - Hypothesis generation Ex: Are single men/women more likely to use coupon than married men/women?
  - Univariate analysis to check the distribution of each feature
    - Histogram
    - Pie Chart
    - Bar Chart
  - Bivariate analysis to check the hypothesis and get insights
    - Multiple bar chart
    - Line chart
  - Correlation and Multi-collinearity

Steps to be followed

- Feature Engineering
  - Derive additional features from one feature
    - Ex: From date you can extract features like month, weekday, day of the month, year, etc.
  - Derive additional features from two or more features
    - Ex: Total price can be calculated based on quantity and per unit price

Steps to be followed

- Apply machine learning model
  - Split the train data in train and test set
  - Train the machine learning model on the train data and calculate the accuracy score over the test data, apply cross validation on entire data
  - Check the feature importance and retrain the model with important/significant feature. You may take input from EDA as well
  - After you are satisfied with the models performance, make predictions on the provided “test set”
  - Create a csv file in the provided format and submit the same.

Suggestions

- Spend more time in understanding the data, obtaining visuals and inferences and rest on model building
- Final presentation should be more business focused than keeping it too technical – Technical details to be added in the appendix

Questions?